On improving the forecast accuracy of the hidden Markov model

Thomas Rooney

Dissertation submitted in partial fulfilment of the requirements for the degree of Master of Commerce

in

The Faculty of Commerce

Department of Actuarial Science

January 2016

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

Declaration

I, Thomas Jeffrey Amhurst Rooney, hereby declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being or is to be submitted for another degree in this or any other University. I empower the University of Cape Town to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signed by candidate

Signature removed

TJA Rooney

15 November 2016

Date

Abstract

The forecast accuracy of a hidden Markov model (HMM) may be low due first, to the measure of forecast accuracy being ignored in the parameter-estimation method and, second, to overfitting caused by the large number of parameters that must be estimated.

A general approach to forecasting is described which aims to resolve these two problems and so improve the forecast accuracy of the HMM. First, the application of extremum estimators to the HMM is proposed. Extremum estimators aim to improve the forecast accuracy of the HMM by minimising an estimate of the forecast error on the observed data. The forecast accuracy is measured by a score function and the use of some general classes of score functions is proposed. This approach contrasts with the standard use of a minus log-likelihood score function. Second, penalised estimation for the HMM is described. The aim of penalised estimation is to reduce overfitting and so increase the forecast accuracy of the HMM. Penalties on both the state-dependent distribution parameters and transition probability matrix are proposed. In addition, a number of cross-validation approaches for tuning the penalty function are investigated.

Empirical assessment of the proposed approach on both simulated and real data demonstrated that, in terms of forecast accuracy, penalised HMMs fitted using extremum estimators generally outperformed unpenalised HMMs fitted using maximum likelihood.

Acknowledgements

Firstly, I would like express my sincere gratitude to my supervisor, Associate Professor Iain L. MacDonald for his guidance, support, patience and motivation for this dissertation.

I would also like to thank my parents for their unconditional support both generally and, in particular, throughout writing this dissertation.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

Contents

Notation and abbreviations

Notation

Symbol	Meaning	Page
α	tuning parameter for penalised estimation	31
C_t	state occupied by Markov chain at time t	3
$\boldsymbol{C}_{1:t}$	$(C_1, C_2, ..., C_t)$	3
CV(α)	cross-validation score when tuning parameter is equal to α	46
$\boldsymbol{\delta}$	stationary distribution of Markov chain	4
D_{KL}	Kullback-Leibler semimetric	42
$F(\cdot)$	is the general symbol for a distribution function. A subscript is included when it is necessary to clarify which random variable the distribution function relates to.	6
$\boldsymbol{\Gamma}$	transition probability matrix of Markov chain	4
$\boldsymbol{\Gamma}_{i\bullet}$	ith row of $\boldsymbol{\Gamma}$	42
$J(\cdot)$	penalty function	30
$\mathcal{L}(\cdot)$	likelihood function	5
$\boldsymbol{\lambda}$	vector or matrix of state-dependent distribution parameters	4
Λ	parameter space of $\boldsymbol{\lambda}$	4
log	logarithm to base e	
L_B	Bregman loss function	19

L_{GPL}	generalised piecewise linear loss function	20
M	state space of C_t	3
$\mathbb{N}$	positive natural numbers	
Ω	sample space of X_t	3
p_i	probability mass or density function in state i	4
$\mathbf{P}(x)$	diagonal matrix with ith diagonal element $p_i(x)$	5
Φ	distribution function of standard normal random variable	
$\mathbb{R}$	real numbers	
$\mathbb{R}_+$	positive real numbers	
$\mathbb{R}_+^0$	non-negative real numbers	
$s(\cdot, \boldsymbol{\theta})$	score function for HMM parametrised by $\boldsymbol{\theta}$	8
$S_0(\boldsymbol{\theta})$	expected score for HMM parametrised by $\boldsymbol{\theta}$	9
$S_T(\cdot, \boldsymbol{\theta})$	estimate of $S_0(\boldsymbol{\theta})$ for HMM parametrised by $\boldsymbol{\theta}$	10
$\boldsymbol{\theta}$	parameter vector of HMM	4
Θ	parameter space of $\boldsymbol{\theta}$	4
$\widehat{\boldsymbol{\theta}}_T$	estimate of $\boldsymbol{\theta}$ using first T observations	10
$\widehat{\boldsymbol{\theta}}_{T,\alpha}$	estimate of $\boldsymbol{\theta}$ using first T observations and tuning parameter value of α	31
X_t	observation at time t	3
$\boldsymbol{X}_{1:t}$	$(X_1, X_2, ..., X_t)$	3
$\mathbb{Z}$	integers	

Finally, all vectors should be interpreted as column vectors.

Abbreviations

CV	cross-validation
DSS	Δ-sequential sampling
DNM	direct numerical minimisation
e.g.	*exempli gratia*
HMM	hidden Markov model
HVB	*hv*-block
i.i.d.	independent, identically distributed
KLD	Kullback-Leibler divergence
LB	last block
LASSO	least absolute shrinkage and selection operator
MAP	maximum *a posteriori*
MLE	maximum likelihood estimate
OEHS	odd-even half-sampling
OOS	out-of-sample
Q-Q	quantile-quantile
t.p.m.	transition probability matrix
VaR	value-at-risk
VB	*v*-block

CHAPTER 1

Introduction

This dissertation is concerned with improving the forecast accuracy of the HMM. HMMs form a widely used class of statistical model, applied where the distribution of the observed values is assumed to depend on the state of some unobserved Markov process. These models have proved useful for general-purpose modelling and forecasting of time-series data, in particular time series of small counts. When using HMMs for forecasting, it is common for the parameter estimation method to be unrelated to the measure of forecast accuracy. In addition, the number of parameters of a basic HMM is quadratic in the number of states of the Markov process; this can result in overfitting even when the number of states is small. These two problems may lead to poor forecast accuracy. The use of HMMs for forecasting in a wide range of fields such as

- finance (Hassan and Nath, 2005),
- seismology (Ebel et al., 2007),
- climatology (Robertson et al., 2004) and
- electricity pricing (González et al., 2005),

makes modifications to HMM which improve the forecast accuracy essential.

This dissertation makes two major proposals. The first is the use of extremum estimators for estimating the parameters of the HMM. Extremum estimators allow the matching of a measure of forecast accuracy with the method of parameter estimation and aim to improve the forecast accuracy of the HMM. This approach generalises the standard approach of maximum likelihood estimation. Extremum estimators are introduced in Chapter 3 and some general measures of forecast accuracy are discussed in Chapter 4.

The second major proposal is the inclusion of a penalty function in the objective function used to estimate the parameters of the HMM. The aim of a penalty function is to reduce overfitting and hence improve the forecast accuracy of the HMM. The application of penalised estimation to HMMs is not new; see, for example, McGibbon et al. (2014). However, our proposed approach is fairly general; the penalty is not tailored to any particular application. Penalised estimation is introduced in Chapter 5 and a cross-validation approach for tuning the size of the penalty is discussed in Chapter 6.

With regard to the the other chapters, the basic HMM is introduced in Chapter 2. Methods for implementing penalised estimation are described in Chapter 7 and some applications of penalised HMMs are presented in Chapter 8. Finally, concluding remarks and suggestions for further research are provided in Chapter 9.

The order of the Chapters 3 to 7 is intended to replicate the general approach taken by a forecaster using HMMs. The analysis presented is fairly general in that the proposed methods may be applied to a wide range of HMMs. Nonetheless, the applications presented in Chapter 8 focus on discrete-valued series, in particular series of small counts as HMMs have shown notable promise in modelling such types of series (Zucchini and MacDonald, 2009). All proofs are given in Appendix B.

CHAPTER 2

Introduction to hidden Markov models

In this chapter a brief introduction to the basic HMM is given. The organisation of the chapter is as follows. In Section 2.1 the mathematics and parametrisation of a basic HMM are introduced. The forms of the likelihood function and forecast distribution for the HMM are then described in Section 2.2.

2.1 The basic HMM

Consider an observed stochastic process $\{X_t\}_{t\in\mathbb{N}}$ on $\Omega \subseteq \mathbb{R}^q$, with the history of the process from time one to time t denoted by $\boldsymbol{X}_{1:t} = (X_1, X_2, ..., X_t)$. To formulate an m-state HMM, suppose there exists an unobserved m-state stochastic process $\{C_t\}_{t\in\mathbb{N}}$ with state-space $M = \{1, 2, ..., m\}$ and history $\boldsymbol{C}_{1:t} = (C_1, C_2, ..., C_t)$, satisfying the following properties:

$$\mathbb{P}(C_t = i \mid \boldsymbol{C}_{1:(t-1)}) = \mathbb{P}(C_t = i \mid C_{t-1}), \quad i \in M, t \in \mathbb{N}\backslash\{1\}, \tag{2.1}$$

$$\mathbb{P}(X_t \in A \mid \boldsymbol{X}_{1:(t-1)}, \boldsymbol{C}_{1:t}) = \mathbb{P}(X_t \in A \mid C_t), \quad A \subseteq \Omega, t \in \mathbb{N}. \tag{2.2}$$

The first property is the Markov property. The second property implies that X_t is dependent only on C_t; that is, the X_t's are conditionally independent given all the C_t's.

The probability mass (or density) function of X_t given that C_t is in state $i \in M$ is denoted by p_i, with support $\Omega_i \subseteq \mathbb{R}^q$. For the purpose of this dissertation, it is assumed that the m p_is are members of the same family of distributions, described by a class of densities of the form $\{p(x|\lambda) : \lambda \in \Lambda\}$, where λ is a real-valued (possibly vector) parameter with parameter space $\Lambda \subseteq \mathbb{R}^d$. Thus the state-dependent distributions share a common support Ω and p_i is described by a parameter λ_i such that $p_i(x) = p(x|\lambda_i)$ for every $x \in \Omega$. The nomenclature for HMMs is often determined by this family of distributions; for example if $p(x|\lambda)$ is a Poisson distribution then the resulting HMM is termed a 'Poisson-HMM'.

The time-homogeneous Markov chain $\{C_t\}_{t\in\mathbb{N}}$ is characterised by an initial distribution $\boldsymbol{\delta}$ and transition probability matrix (t.p.m.) $\boldsymbol{\Gamma} = (\gamma_{ij})$ where $\gamma_{ij} \in [0, 1]$. A common simplifying assumption is that $\boldsymbol{\delta}$ be taken as the stationary distribution for $\boldsymbol{\Gamma}$; that is, $\boldsymbol{\delta}$ satisfies $\boldsymbol{\delta}'\boldsymbol{\Gamma} = \boldsymbol{\delta}'$ in which the case the Markov chain is said to be stationary. Equivalently, stationarity is imposed by assuming the observable and hidden stochastic processes are indexed by the integers, in which case they are said to be 'doubly infinite' (Leroux, 1989). In this case we write $\{X_t\}_{t\in\mathbb{Z}}$ and $\{C_t\}_{t\in\mathbb{Z}}$ with suitable modifications of Equation 2.1 and Equation 2.2. If in addition the Markov chain is assumed ergodic then the stationary distribution is unique and determined by the t.p.m. $\boldsymbol{\Gamma}$. For the purpose of this work, it shall be assumed that the hidden Markov chain is indeed ergodic and that the hidden and observed processes are treated as doubly infinite; the reasons for this are described later.

Hence a stationary m-state HMM with a set of specified state-dependent distributions is specified by a matrix of parameters for the state-dependent distributions,

$$\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1' \\ \vdots \\ \lambda_m' \end{pmatrix}$$

and a t.p.m. $\boldsymbol{\Gamma}$. For notational ease, we define a parameter vector $\boldsymbol{\theta} \in \Theta$,

where Θ is a subset of a Euclidean space, such that

$$\boldsymbol{\theta} = (\gamma_{11}, \gamma_{12}, ..., \gamma_{mm}, \lambda_1', ..., \lambda_m'),$$

which we term the parameter vector of the HMM.

2.2 Likelihood functions and forecast distributions

The likelihood of the observations [1] $\boldsymbol{x}_{1:T} = (x_1, ..., x_T)$, assumed to be generated from a m-state HMM has a convenient and explicit form. Let $\mathbf{P}(x)$ be a real diagonal matrix with the ith diagonal entry equal to $p_i(x)$, then Zucchini and Guttorp (1991) showed the likelihood of $\boldsymbol{x}_{1:T}$ is equal to

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_{1:T}) = \boldsymbol{\delta}'\boldsymbol{\Gamma}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)...\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1} = \boldsymbol{\delta}'\left(\prod_{t=1}^{T}\boldsymbol{\Gamma}\mathbf{P}(x_t)\right)\mathbf{1}, \qquad (2.3)$$

where $\mathbf{1}$ is a $m \times 1$ vector of ones. The above expression requires $\mathcal{O}(Tm^2)$ calculations; feasible even for large values of T.

Another convenient property of the likelihood is the ease with which missing data are accommodated; if observation x_t is missing then the associated $\mathbf{P}(x_t)$ in the likelihood is simply replaced by the identity matrix. This property also ensures that no inconsistencies are created regarding the start point of the observation vector. This statement is best explained by an example: begin by observing $\boldsymbol{x}_{1:T}$. Then suppose the observation window is extended by starting at some point k, a non-positive integer, and the observations $(x_k, x_{k+1}, ..., x_0)$ are treated as missing. Then the likelihood is

$$\mathcal{L}(\boldsymbol{\theta}, (x_k, ..., x_0, \boldsymbol{x}_{1:T})) = \boldsymbol{\delta}'\boldsymbol{\Gamma}^{|k|+1}\left(\prod_{t=1}^{T}\boldsymbol{\Gamma}\mathbf{P}(x_t)\right)\mathbf{1} = \boldsymbol{\delta}'\left(\prod_{t=1}^{T}\boldsymbol{\Gamma}\mathbf{P}(x_t)\right)\mathbf{1},$$

due to stationarity. The above term is just $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_{1:T})$, demonstrating the likelihood is invariant to 'artificial' modifications to the beginning of the

[1] As $\{X_t\}_{t\in\mathbb{Z}}$ is doubly infinite $t = 1$ should be interpreted as the time when observation of the process begins.

observation period. This result may seem obvious but will come in use later in the dissertation.

Before stating the h-step forecast distribution, it is notationally convenient to introduce the vector $\boldsymbol{\alpha}_t$ for $t \in \mathbb{N}$ where

$$\boldsymbol{\alpha}_t' = \boldsymbol{\delta}' \left(\prod_{k=1}^{t} \boldsymbol{\Gamma} \mathbf{P}(x_k) \right), \tag{2.4}$$

where, by convention, $\boldsymbol{\alpha}_0' = \boldsymbol{\delta}'$. Note that $\boldsymbol{\alpha}_t' \mathbf{1}$ is equal to $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_{1:t})$. The h-step forecast density (or mass) function is then given by

$$f_{t+h,\boldsymbol{\theta}}(x) = \frac{\boldsymbol{\alpha}_t' \boldsymbol{\Gamma}^h \mathbf{P}(x) \mathbf{1}}{\boldsymbol{\alpha}_t' \mathbf{1}}. \tag{2.5}$$

Note that this term is similarly invariant to the extension of the observation window described above. The associated forecast distribution is denoted by

$$F_{t+h,\boldsymbol{\theta}}(x) = \mathbb{P}_{\boldsymbol{\theta}}(X_{t+h} \leq x).$$

Of primary interest in this dissertation is the case for $h = 1$. Thus for brevity by 'forecast distribution' without a qualifier is meant the 1-step forecast distribution and similarly for the forecast density (or mass) functions.

CHAPTER 3

Forecasting and extremum estimators

One major aim of time series analysis is to make useful forecasts of observable phenomena. The focus of this chapter is on forecasting with HMMs and estimation of the parameters of an HMM. This chapter introduces a general framework for forecasting, critical to which is a measure of the forecast accuracy. A class of estimators for the HMM, termed extremum estimators, is then proposed which links the estimation procedure with the measure of forecast accuracy. This is a more general approach than the standard method of maximising the likelihood; see Rabiner (1989), Cappé et al. (2005) and Zucchini and MacDonald (2009).

The chapter proceeds as follows. In Section 3.1 the basic approach to forecasting is introduced. In Section 3.2 a method of parameter estimation for forecasting with HMMs is proposed. Section 3.3 covers the consistency of these parameter estimates. Finally, in Section 3.4 some methods for model checking are discussed.

3.1 The basic forecasting approach

Term 'forecasting' as used here refers to the process of issuing a statistical statement at time t about a set of random variables $\{X_{t+1}, X_{t+2}, ..., X_{t+h}\}$ which are not yet observed. The value of h is the forecast window; the period over which a forecast is made which, in principle, should be determined by the context in which a forecast is required. However, for the general purpose proposed here consideration will be given only to the case $h = 1$, which is termed a one-step-ahead forecast. There are three reasons for this. First, one-step-ahead forecasting is fairly common, the four applications mentioned above using one-step-ahead forecasting. Second, there are fewer asymptotic and convergence results for estimators in the case $h > 1$. Third, the period of forecast is not a focus of this work; the simplest approach of taking $h = 1$ does not detract from the analysis.

To make the meaning one-step-ahead forecasting precise, suppose at some time t a forecaster with access to the history of the observed process, $\boldsymbol{x}_{1:t}$, wants to make a statistical statement about X_{t+1}. The approach of Dawid (1984), termed 'probabilistic forecasting', is for the forecaster to issue a cumulative forecast distribution $F_{t+1,\boldsymbol{\theta}}$. In the present context, $F_{t+1,\boldsymbol{\theta}}$ is the forecast distribution implied by an HMM with parameter $\boldsymbol{\theta}$ and history of the process $\boldsymbol{x}_{1:t}$. The usefulness of probabilistic forecasting is twofold. First, the forecast distribution provides a means by which a point forecast can be made, for example by taking expectation of X_{t+1} with respect to $F_{t+1,\boldsymbol{\theta}}$. Second the forecast distributions provide a measure of the uncertainty for that forecast.

The forecaster must provide in addition a score function[1]

$$s : \Omega \times \Theta \longrightarrow \mathbb{R},$$

which compares a forecast distribution $F_{t+1,\boldsymbol{\theta}}$ and the realised value of X_{t+1},

[1]To avoid confusion, for the purposes of this dissertation 'score' should always be interpreted as defined here. In particular, score is not used to refer to the derivative of the log-likelihood with respect to the parameter vector.

x_{t+1}, in order to measure the accuracy of a forecast or compare the accuracies of two different forecasts. We shall assume that the score function is negatively-orientated, in which case it is interpreted as a loss function. There is no canonical choice of score function in time-series analysis. A wide variety of measures are proposed in the literature; Hyndman and Koehler (2006) provide a summary. We will follow the approach of Gneiting (2011) by assuming that the score function measures the loss accrued by the forecaster, which links closely to the Bayesian interpretation of the score function as negative utility. The usefulness of this approach, provided the score function represents accurately the loss to the forecaster, is the implication that the forecaster must want to minimise the score. Put differently, the score function should be specified with respect to a particular forecaster and problem and the forecasting procedure should proceed under the assumption that the forecaster wants to minimise the score; it is not in the hands of the model builder to determine the score function. For the sake of brevity, 'score' is henceforth to mean 'forecast score' or 'score of a forecast'.

3.2 Extremum estimators for the HMM

Having specified a particular score function s, the next step is to find an estimate $\widehat{\boldsymbol{\theta}}_T \in \Theta$ for $\boldsymbol{\theta}$ which, in some sense, minimises the score of the forecast. The qualifier 'in some sense' accounts for the contradiction inherent in minimising the score. The nature of a forecast implies uncertainty about a quantity required to calculate the score and thus uncertainty about the score itself, the minimisation of which is not precisely defined. Hence a method of transforming the distribution for the score into a single real number is required. The approach we follow is to define $S_0(\boldsymbol{\theta})$ where

$$S_0(\boldsymbol{\theta}) = \mathbb{E}[s(X_t, \boldsymbol{\theta})] = \int_\Omega s(x, \boldsymbol{\theta}) \, \mathrm{d}F_t(x); \tag{3.1}$$

that is, $S_0(\boldsymbol{\theta})$ is the expected score for a given value of $\boldsymbol{\theta}$. The actual generating distribution function of X_t is denoted by F_t; we emphasise the difference

between this distribution function and the forecast distribution implied by an HMM with parameter vector $\boldsymbol{\theta}$; denoted $F_{t,\boldsymbol{\theta}}$. We proceed under the assumption that the forecaster wishes to minimise the expected score.

$S_0(\boldsymbol{\theta})$ is never normally observed as F_t is not known; at best an approximating function for $S_0(\boldsymbol{\theta})$ is found and minimised over $\boldsymbol{\theta}$. This approach generalises 'empirical risk minimisation'; a learning principle for independent and identically (i.i.d.) data where the risk is approximated with a standard Monte Carlo average (Vapnik, 2000).

We now present a general approach for estimation of $\boldsymbol{\theta}$. We propose use of wide class of parametric estimators termed 'extremum estimators'; see Amemiya (1985). Despite their great generality, extremum estimators exhibit a number of useful properties, and are thus useful for a general approach to estimation; a detailed description of these properties is provided in Section A.2. Formally, $\widehat{\boldsymbol{\theta}}_T$ is said to be an extremum estimator if an objective function $S_T : \Omega^T \times \Theta \longrightarrow \mathbb{R}$ exists such that

$$\widehat{\boldsymbol{\theta}}_T \in \arg\min_{\boldsymbol{\theta}\in\Theta} S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta}). \tag{3.2}$$

This broad class of estimators includes maximum likelihood, generalised least squares and generalised methods of moments (Hansen, 1982). The body of study of extremum estimators is primarily from an econometric perspective; see, for example, Hayashi (2000). For the purposes of this dissertation, the class of objective functions is restricted to those of M-estimator (Huber, 2011) form:

$$S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} s(\boldsymbol{x}_t, \boldsymbol{\theta}),$$

where s is the score function specified by the forecaster. This is the approach of Singleton (2009), which offers great intuitive appeal; the parameters are estimated by minimising an average of scores on the dataset. Indeed, a set of sufficient conditions are presented in Appendix A which ensure that $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ tends to $\mathbb{E}[s(X_t, \boldsymbol{\theta})]$ as t tends to infinity.

3.3 Consistency of extremum estimators

Of course, the usefulness of the extremum estimator S_T is that if $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ can be regarded as an approximating function for $\mathbb{E}[s(X_t, \boldsymbol{\theta})]$, then a value of $\boldsymbol{\theta}$ minimising $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ should approximate well a value of $\boldsymbol{\theta}$ minimising $\mathbb{E}[s(X_t, \boldsymbol{\theta})]$[2]. Thus it is critical to show that a value of $\boldsymbol{\theta}$ minimising $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$, that is $\widehat{\boldsymbol{\theta}}_T$, will in some sense tend to a value of $\boldsymbol{\theta}$ minimising $S_0(\boldsymbol{\theta}) = \mathbb{E}[s(X_t, \boldsymbol{\theta})]$. This property is termed 'consistency', a concept studied extensively in the asymptotic theory of statistics; see, for example, DasGupta (2008).

However, although this is a critical property, it is not of direct relevance to the aims of this dissertation. Thus a proof of the consistency of the extremum estimator for the HMM is left to Appendix A.

3.4 Model checking and measuring uncertainty of a forecast

Thus far it has be assumed that the forecaster wants only to minimise the expected score. This is a convenient assumption for deriving the extremum estimator but is ultimately unrealistic. The main aim of the forecaster is to provide useful forecasts; even if the objective of minimising the expected score is met, the resulting model may not provide useful forecasts. For the purposes of this dissertation, a useful forecasting model is one which can provide accurate forecasts at an acceptable level of certainty. Pseudo-residuals will be used to assess the accuracy of the model, and forecast intervals used to measure the uncertainty of a forecast.

[2]It is assumed that $\mathbb{E}[s(X_t, \boldsymbol{\theta})]$ exists and is finite.

3.4.1 Checking a forecasting model with pseudo-residuals

To assess the adequacy of a forecast model, use is made of forecast pseudo-residuals, which were proposed by Zucchini and MacDonald (2009) and follow the same principle as a technique termed 'probability integral transform values' (Dawid, 1984).

The basic technique of forecast pseudo-residuals is as follows. Suppose X_t is a univariate continuous random variable with distribution function F_t. Then $F_t(X_t)$ is uniformly distributed on the unit interval. Denote by $\boldsymbol{x}_{1:T}$ a series of observations of X_t and let $F_{t,\boldsymbol{\theta}}$ denote the forecast distribution for X_t. Let

$$u_t = F_{t,\boldsymbol{\theta}}(x_t),$$

which we term the uniform pseudo-residual of x_t. Since the $F_t(X_t)$ are i.i.d. $U(0,1)$, it follows that if the $F_{t,\boldsymbol{\theta}}$s are good approximations of the F_ts, so the empirical distribution of the u_ts should be approximately uniform. Thus by examining the empirical distribution of the u_ts the model can be checked; if the empirical distribution of the u_ts is not close to uniform then this may indicate the model is not valid. To avoid problems with the visual analysis of uniform pseudo-residuals, Zucchini and MacDonald (2009) suggest transforming each uniform pseudo-residual to a normal pseudo-residual, given by

$$z_t = \Phi^{-1}(u_t),$$

which are distributed standard normal if the model is valid; a Q-Q plot (Wilk and Gnanadesikan, 1968) is a useful visual aid for analysis of these residuals.

If X_t is discrete, the pseudo-residuals are defined as intervals. More precisely, the uniform pseudo-residual segments are defined as

$$[u_t^-; u_t^+] = [F_{t,\boldsymbol{\theta}}(x_t^-), F_{t,\boldsymbol{\theta}}(x_t)],$$

where x_t^- denotes the greatest possible realisation of X_t that is strictly less than x_t, and the normal pseudo-residual segments are defined as

$$[z_t^-; z_t^+] = [\Phi^{-1}(u_t^-), \Phi^{-1}(u_t^+)].$$

In order to check for normality, a single value as opposed to interval is required. In this case, we make use of normal randomised quantile-residuals (Dunn and Smyth, 1996) which are defined as

$$z_t^m = \Phi^{-1}\left(u_t^m\right),$$

where u_t^m is a sample from a uniformly distributed random variable on the interval (u_t^-, u_t^+).

If the observed series is multivariate, then we follow the approach of Diebold et al. (1998) which is as follows. Suppose that $\boldsymbol{X}_t = (X_{t1}, X_{t2}, ..., X_{tq})$ is a vector of length q. Let $F_{t,\boldsymbol{\theta}}^k$ denote the distribution function for X_{tk} given $(\boldsymbol{X}_t)_{1:(k-1)}$; that is,

$$F_{t,\boldsymbol{\theta}}^k(x) = \mathbb{P}_{\boldsymbol{\theta}}\left(X_{tk} \leq x \mid X_{t,k-1}, ..., X_{t1}\right).$$

By convention, $F_{t,\boldsymbol{\theta}}^1$ is the marginal distribution function for X_{t1}. If it is assumed that $F_{t,\boldsymbol{\theta}}^k$ is the actual distribution function of $X_{tk}|(\boldsymbol{X}_t)_{1:(k-1)}$, then it can then be shown that the $\{F_{t,\boldsymbol{\theta}}^k(X_{tk})\}_{t=1}^T$ are i.i.d. $U(0,1)$ and, furthermore, that $F_{t,\boldsymbol{\theta}}^k(X_{tk})$ is independent of $F_{s,\boldsymbol{\theta}}^r(X_{sr})$ for all $t, s \in \{1, ..., T\}$ and $k, r \in \{1, ..., q\}$. Thus, for every value of t and k we find the pseudo-residuals

$$u_{t,k} = F_{t,\boldsymbol{\theta}}^k(x_{tk}),$$

which are approximately i.i.d. $U(0,1)$ if the HMM parameterised by $\boldsymbol{\theta}$ is a good fit to the data.

Finally, we emphasise the distinction between model checking and minimising $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$; the latter is useful only as a relative measure of forecast accuracy between candidate models. In contrast, the purpose of model checking is to assess the forecast accuracy on an absolute level.

3.4.2 Forecast intervals and potential risk

Suppose that a forecaster wishes to issue a forecast about X_{t+1} given $\boldsymbol{x}_{1:t}$. Intuitively, uncertainty about X_{t+1} can be thought to arise from two sources.

The first is the uncertainty about X_{t+1} given the actual generating distribution for X_{t+1}, F_{t+1}. The second source of uncertainty is the difference between the actual generating distribution and the forecast distribution, $F_{t+1,\boldsymbol{\theta}}$. The purpose of using pseudo-residuals is to check the fit of the forecast distribution to the actual generating distribution; nothing is said by the pseudo-residuals of uncertainty resulting from the forecast distribution. The key point is that even if $F_{t+1} = F_{t+1,\boldsymbol{\theta}}$, the uncertainty arising from F_{t+1} may be too high to allow useful forecasting.

Thus far the term 'uncertainty' has not been precisely defined; to proceed we require some measure of the uncertainty arising from F_{t+1}. One approach in time-series analysis is to issue a '$100(1-\alpha)\%$ forecast region'; a subset A of Ω such that X_{t+1} has $100(1-\alpha)\%$ probability of being in A. Of course, as F_{t+1} is unknown, we must use our approximation to this distribution; that is, $F_{t+1,\boldsymbol{\theta}}$. In the univariate case this region is an interval, denoted (l, u) where $l, u \in \Omega$ are such that

$$\mathbb{P}_{\boldsymbol{\theta}}(X_{t+1} \leq l) = \mathbb{P}_{\boldsymbol{\theta}}(X_{t+1} \geq u) = \alpha/2.$$

The resulting interval is termed a $100(1-\alpha)\%$ forecast interval, where $\alpha \in (0, 1)$. Clearly

$$\mathbb{P}_{\boldsymbol{\theta}}(l \leq X_{t+1} \leq u) = 1 - \alpha.$$

There are, however, two problems with forecast intervals. First, the multivariate case is fairly complex and, second, such intervals may not be particularly useful. As stated earlier, the forecaster is directly concerned with the expected score only; the actual forecast value is only important in so far as it determines the score of that forecast. The weakness of forecast intervals is that focus is given to uncertainty of the forecast, but nothing is said about the implications of this uncertainty for the potential score. Hence, as an alternative to forecast intervals, we introduce a measure which we term 'potential risk'. If X_t is continuous we define the potential risk as the value r such that

$$\mathbb{P}_{\boldsymbol{\theta}}(s(X_{t+1}, \boldsymbol{\theta}) \leq r) = 1 - \alpha.$$

The potential risk may be interpreted as the maximum possible score at a given confidence level and is analogous to the 'value at risk' measure used in finance; see Duffie and Pan (1997). As an additional advantage, potential risk does not suffer the problems faced by forecast intervals when X_{t+1} is multivariate.

If X_t is discrete then $s(X_t, \boldsymbol{\theta})$ is also discrete. In this case, we take an approach analogous to that taken with the pseudo-residuals by defining the potential risk as the interval $[r_-, r_+]$ such that r_- is the largest possible value of $s(X_t, \boldsymbol{\theta})$ satisfying

$$\mathbb{P}_{\boldsymbol{\theta}}(s(X_{t+1}, \boldsymbol{\theta}) \leq r_-) \leq 1 - \alpha. \tag{3.3}$$

Similarly, r_+ is the smallest possible value of $s(X_t, \boldsymbol{\theta})$ satisfying

$$\mathbb{P}_{\boldsymbol{\theta}}(s(X_{t+1}, \boldsymbol{\theta}) \leq r_+) \geq 1 - \alpha. \tag{3.4}$$

We emphasise that the potential risk is not a measure of goodness-of-fit; it describes a probabilistic bound on the score by assuming that the estimated parameters are the 'true' parameters. If the estimated parameters are a good fit to the true parameters, then the estimated potential risk will be close to the true potential risk. Furthermore, the forecaster should *ceteris paribus* prefer the model with lower potential risk. Examples of how potential risk is used are given in Chapter 8.

CHAPTER 4

Two general forms of score functions

Thus far little has been said of the actual form of the score function. This chapter remedies this problem by describing the minus log-likelihood score function and point-forecast based score functions. The forms presented are fairly broad; the purpose being to encompass a wide range of practically relevant score functions. For each score form, some general theory is discussed as well as the practical relevance of each score function.

4.1 The minus log-likelihood score function

An important example of an extremum estimator is the maximum likelihood estimator. Parameter estimation for the HMM by maximisation of the likelihood function is a standard approach; see Rabiner (1989), Cappé et al. (2005) and Zucchini and MacDonald (2009). Maximum likelihood estimation is placed in the context of extremum estimators by supposing that the score function is the minus log-likelihood for x_t, that is,

$$\begin{aligned} s(x_t, \boldsymbol{\theta}) &= -\log \mathbb{P}\left(X_t = x_t \mid \boldsymbol{X}_{1:(t-1)} = \boldsymbol{x}_{1:(t-1)}\right) \\ &= -\log f_{t,\boldsymbol{\theta}}(x_t). \end{aligned}$$

Here $f_{t,\boldsymbol{\theta}}$ is the forecast density (or mass) function for X_t given the history of $\{X_t\}_{t\in\mathbb{Z}}$ from time one to $t-1$, the density function being that arising from an HMM with parameter vector $\boldsymbol{\theta}$. A parameter estimate is thus

$$\widehat{\boldsymbol{\theta}}_T \in \arg\min_{\boldsymbol{\theta}\in\Theta} -\frac{1}{T}\sum_{t=1}^{T} \log f_{t,\boldsymbol{\theta}}(x_t) \tag{4.1}$$

In this case $\widehat{\boldsymbol{\theta}}_T$ is the usual maximum likelihood estimate (MLE), the standard method of estimation for HMMs. Put differently, a score function of the form above is implied by estimating the parameters of an HMM using maximum likelihood.

4.2 Point-forecast based score functions

The minus log-likelihood score function compares the forecast distribution in its entirety to a realised value. Therefore the forecast issued is a distribution of values as opposed to a single point from that distribution. However, in practical application, decision making often requires a single point forecast to be made and the loss incurred to be a function of the forecast value and realised value. Economics provides many examples of where such estimates are required, be it the price of a share tomorrow in order to value a derivative, or the demand for a particular product in a year in order to determine production levels. In these cases a decision regarding a single point forecast is required and the loss function has an easily used financial interpretation. More generally, we term score functions for this purpose 'point-forecast based score functions' and describe in this section two general classes of such score functions: Bregman and generalised piecewise linear loss functions. A great appeal of issuing a point forecast over a forecast distribution is that forecasters should find it much easier to describe their losses. Thus, point-forecast based score functions may be more easily interpreted than their likelihood-based counterpart and hence have more practical use.

4.2.1 Pairwise loss functions and optimal point predictors

This section follows the approach of Gneiting (2011). Assume again that the forecaster has at time t access to a cumulative forecast distribution $F_{t+1,\boldsymbol{\theta}}$ for x_{t+1} from the HMM. The forecaster must then provide a point forecast $\widehat{x}_{t+1}$ based upon this forecast distribution. Formally, define a statistical functional $\mathcal{D} : \mathcal{F} \longrightarrow \Omega$, which maps a forecast distribution for x_{t+1} into a point forecast $\widehat{x}_{t+1} \in \mathcal{D}(F_{t+1,\boldsymbol{\theta}})$. The set notation allows for generality; in many cases $\mathcal{D}(F_{t+1,\boldsymbol{\theta}})$ is a singleton but this is not true in general; for example, if $\mathcal{D}(F_{t+1,\boldsymbol{\theta}})$ is the mode of $F_{t+1,\boldsymbol{\theta}}$, and $F_{t+1,\boldsymbol{\theta}}$ has two global maxima. The score function should then be a function of the forecast and realised values; that is,

$$s(x_{t+1}, \boldsymbol{\theta}) = L\left(x_{t+1}, \mathcal{D}(F_{t+1,\boldsymbol{\theta}})\right) \tag{4.2}$$

where $L : \Omega \times \Omega \longrightarrow \mathbb{R}_0^+$ is a suitably defined loss function; suitable here meaning an accurate measure of the forecaster's loss. The loss function will be provided by the forecaster so a choice of statistical functional $\mathcal{D}$ must still be made. This choice is important; clearly the form of $\mathcal{D}$ will in general affect the forecast made and hence the score. We assume, as earlier, that the forecaster wants to pick $\mathcal{D}$ in a way which minimises the expected score for a particular forecast distribution. More formally, and noting the score is just a multiple of the loss function, the approach is to pick $\mathcal{D}$ so that $\widehat{x}_{t+1} \in \mathcal{D}(F_{t+1,\boldsymbol{\theta}})$ only if

$$\widehat{x}_{t+1} \in \arg\min_{x \in \Omega} \mathbb{E}[L(X_{t+1}, x)],$$

where the expectation is taken over X_{t+1}, which has cumulative distribution $F_{t+1,\boldsymbol{\theta}}$. Thus, for any particular loss function L, $\mathcal{D}$ is chosen to minimise the expected loss. Gneiting terms functionals meeting the aforementioned property as 'optimal point predictors'.

Unfortunately the description of loss functions and optimal point predictors in the multivariate case is complex. We therefore suppose for the

moment that $\Omega \subseteq \mathbb{R}$. An aim of this section, being consistent with an aim of this dissertation, is to consider a set of loss functions that are relevant and general enough for use in a wide range of applications. Two common forms for loss functions are the quadratic form, $L(x, y) = (x - y)^2$, and the linear form, $L(x, y) = |x - y|$. Both of these loss functions are symmetric, in that $L(x, y) = L(y, x)$, and of the prediction error form, in that the loss depends on the difference between x and y only. However, there is a substantial amount of literature (see Gneiting (2008) for a review) indicating that practically relevant loss functions are neither symmetric nor of the prediction error form. Thus we begin by considering two classes of univariate loss functions which are in general asymmetric and not of the prediction error form. We later generalise these loss functions to the multivariate case.

4.2.2 Univariate Bregman loss functions

The first class we consider are the Bregman loss functions. Before introducing this class, it is necessary to define the notion of a subgradient due to Rockafellar (1970): if ϕ is a mapping from Ω to $\mathbb{R}$ then a real-valued function ψ on Ω is a subgradient of ϕ if

$$\phi(y) \geq \phi(x) + \psi(x)(y - x) \text{ for all } x, y \in \Omega.$$

If ϕ is differentiable at some point x in the interior of Ω then its subgradient $\psi(x)$ is unique and equals the derivative at x (Gneiting, 2008). Thus the subgradient generalises the derivative of a function for certain non-differentiable functions, a particularly important case being $\phi(x) = |x|$. The Bregman loss functions (Banerjee et al., 2005) are then of the form

$$L_{\mathrm{B}}(x, y) = \phi(y) - \phi(x) - \psi(y)(x - y),$$

where $\phi : \Omega \longrightarrow \mathbb{R}$ is a convex function with subgradient ψ. Note that $L_{\mathrm{B}}(x, y)$ is continuous and non-negative, by definition of the subgradient. A particularly important subclass termed 'power loss functions ' is derived by

taking $\phi(x) = |x|^a$ where $a > 1$. The resulting loss function is

$$L_{\mathrm{B}}(x, y) = |x|^a - |y|^a - a\,\mathrm{sign}(y)|y|^{a-1}(x - y), \tag{4.3}$$

which is equal to quadratic loss when $a = 2$. An important property of Bregman loss functions is given in the following theorem.

Theorem 1. *Suppose that the loss function is of Bregman form and the technical conditions* $\mathbb{E}_{\boldsymbol{\theta}}[X_t] < \infty$ *and* $\mathbb{E}_{\boldsymbol{\theta}}[\phi(X_t)] < \infty$ *are met. Then the expected value of* X_t *with respect to the forecast distribution,*

$$\mathcal{D}(F_{t,\boldsymbol{\theta}}) = \int_{\Omega} x \, dF_{t,\boldsymbol{\theta}}(x),$$

is an optimal point predictor.

In fact, the converse of Theorem 1 is true. It is possible to show, subject to some conditions on the loss function, that choosing the expected value functional for $\mathcal{D}$ and requiring the expected loss to be minimised implies that the loss function must be of Bregman form. The reader is directed to Gneiting (2008) for further details. Taking the point forecast as the expected value of the random variable being forecast is common (Gneiting, 2011) which makes loss functions of the Bregman form an important area of study.

4.2.3 Univariate generalised piecewise linear loss functions

The second class of point-forecast score function we consider are generalised piecewise linear (GPL) loss functions (Schlaifer and Raiffa, 1961), of the form

$$L_{\mathrm{GPL}}(x, y) = \begin{cases} (1 - \beta)(g(y) - g(x)) & \text{if } x \leq y \\ \beta(g(x) - g(y)) & \text{if } x \geq y \end{cases}$$

where $g : \Omega \longrightarrow \mathbb{R}_0^+$ is a non-decreasing function on Ω and β is an 'order parameter' in $(0, 1)$. This class of loss of functions offers great intuitive appeal; g can be thought of as a utility function for the future quantity, elicitable

in many practical situations so that the loss depends on the difference in utility of the realised and forecast value. Following from Equation 4.2, x is the realised value and y the forecast value. Thus in the case of an overprediction ($x < y$) the loss is proportional to $g(y) - g(x)$ and, in the case of underprediction ($y < x$), proportional to $g(x) - g(y)$. The key is that the loss is asymmetric; high values of β imply greater relative aversion to underprediction than to overprediction. The opposite is true for small values of β. For the purposes of this dissertation it is required that g is differentiable, which implies almost everywhere differentiability of $L_{\text{GPL}}(x, y)$ with respect to x and y (Gneiting, 2008).

The optimal point predictor is specified by the following theorem.

Theorem 2. *Suppose that the loss function is GPL of order β and the technical condition $\mathbb{E}_{F_{t,\boldsymbol{\theta}}}[g(X_t)] < \infty$ is met. Then the β-quantile of the forecast distribution,*

$$\mathcal{D}(F_{t,\boldsymbol{\theta}}) = F^{-1}_{t,\boldsymbol{\theta}}(\beta),$$

is an optimal point predictor.

Theorem 2 highlights a very powerful property of GPL loss functions. That property is that the optimal point predictor does not depend on the form of the utility function g. Indeed, provided β is known, the forecaster can issue an optimal point forecast with the only known fact about g being that it is non-decreasing. This is particularly useful as often g is unknown or itself estimated under uncertainty. The result of Theorem 2 is also intuitive: a value of $\beta > 0.5$ indicates greater relative aversion to underprediction than to overprediction, so it appears to be prudent to forecast a value which is more likely to be an overprediction than an underprediction; that is, to use a quantile above the median of the forecast distribution.

Analogous to Bregman loss functions, it is possible to show, subject to some conditions on the loss function, that choosing the β-quantile for $\mathcal{D}$ and requiring the expected loss to be minimised implies that the loss function must be of GPL form (Gneiting, 2008).

4.2.4 Multivariate extensions

We now extend the discussion to general Euclidean sample spaces; that is, $\Omega \subseteq \mathbb{R}^q$. Bregman loss functions and Theorem 1 generalise well to the multivariate case; the component-wise expectation[1] is an optimal point predictor for multivariate Bregman loss functions of the form

$$L_{\mathrm{B}}(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \nabla\phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle,$$

where $\phi : \mathbb{R}^q \longrightarrow \mathbb{R}$ is convex with gradient $\nabla\phi : \mathbb{R}^q \longrightarrow \mathbb{R}^q$, and $\langle \quad , \quad \rangle$ denotes the scalar product (Banerjee et al., 2005). This result holds subject to one of two smoothness conditions on ϕ. One condition due to Osband and Reichelstein (1985) is to require continuity of the derivative of $L_{\mathrm{B}}(\boldsymbol{x}, \boldsymbol{y})$ with respect to $\boldsymbol{y}$. Another condition is to require continuity of the second derivative of $L_{\mathrm{B}}(\boldsymbol{x}, \boldsymbol{y})$ with respect to $\boldsymbol{x}$ (Banerjee et al., 2005).

In contrast to the Bregman loss functions, the multivariate generalisation of GPL loss functions and Theorem 2 is complex (Gneiting, 2011). Certainly, there is appeal in again supposing the existence of a utility-like function $g : \Omega \longrightarrow \mathbb{R}_0^+$ which maps the vector outcome into a single value. A natural extension to a multivariate GPL is then given by

$$L_{\mathrm{GPL}}(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} (1-\beta)(g(\boldsymbol{y}) - g(\boldsymbol{x})) & \text{if } g(\boldsymbol{x}) \leq g(\boldsymbol{y}) \\ \beta(g(\boldsymbol{x}) - g(\boldsymbol{y})) & \text{if } g(\boldsymbol{x}) \geq g(\boldsymbol{y}) \end{cases}$$

This loss function implies no analytical form for the optimal point predictor; rather it must be found numerically. This approach may not be feasible for the purposes of this dissertation, as each evaluation of $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ would require T separate minimisations. As an alternative, one can start with taking the quantile function of multivariate distributions as the optimal point predictor, and then develop a consistent loss function. Unfortunately, such

[1]For a random variable $\boldsymbol{X} = (X_1, ..., X_q) \in \mathbb{R}^q$, by component-wise expectation is meant the vector of component expectations $\mathbb{E}[\boldsymbol{X}] = (\mathbb{E}[X_1], ..., \mathbb{E}[X_q])$ where each expectation is taken with respect to the marginal distribution implied by the forecast distribution for $\boldsymbol{X}$.

quantile functions are not easy to define; Serfling (2002) cites five different approaches to defining multivariate quantile functions.

We develop a simple multivariate loss function based upon quantiles of Abdous and Theodorescu (1992) form. Thus consider a univariate GPL loss function with g the identity function. It easily shown that the loss function is equal to

$$\frac{|x-y|+(2\beta-1)(x-y)}{2}.$$

Suppose now that $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^q$ for $q > 1$. A natural elementwise generalisation of this loss function would be to apply univariate GPL loss elementwise across $\boldsymbol{x}$ and $\boldsymbol{y}$, with the resulting vector mapped to a real number by some L_p norm. Thus for $p \in [1, \infty)$ and $\beta \in (0, 1)$, a norm-like functions $\|\boldsymbol{x}\|_{p,\beta}$ is defined as

$$\|\boldsymbol{x}\|_{p,\beta} = \left\| \frac{|x_1|+(2\beta-1)x_1}{2}, ..., \frac{|x_q|+(2\beta-1)x_q}{2} \right\|_p.$$

Then define the multivariate piecewise loss function as

$$L_{\mathrm{PL}}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_{p,\beta},$$

which generalises GPL loss with g the identity function. Under this loss function and provided $|\mathbb{E}[\boldsymbol{X}_t]| < \infty$, an optimal point estimate is of the form

$$\widehat{\boldsymbol{x}_t} \in \underset{\boldsymbol{x}\in\Omega}{\arg\min}\, \mathbb{E}[\|\boldsymbol{X}_t - \boldsymbol{x}\|_{p,\beta}]. \tag{4.4}$$

However, Equation 4.4 is the exact definition of the β-quantile given by Abdous and Theodorescu (1992), for some value of p. The importance of p should not be overlooked; it can be set to reflect accurately the forecaster's loss, and will determine the optimal point forecast. Indeed, under quantiles of the Abdous and Theodorescu form, it is meaningless to state a β-quantile without providing a value of p. For $p = 1$, the expected value in Equation 4.4 splits into the form

$$\sum_{i=1}^{q} \mathbb{E}\left[\frac{|X_{ti}-x_i|+(2\beta-1)(X_{ti}-x_i)}{2}\right],$$

which can be minimised elementwise with x_i equal to the β-quantile of the marginal forecast distribution for X_{ti}. For $p = 2$ and $\beta = 0.5$, the well studied 'spatial median' is obtained; see, for example, Small (1990). In general, for $p > 1$, there is no analytical solution to Equation 4.4. Thus Abdous and Theodorescu (1992) propose a subgradient method (see, for example, Polyak (1987)) for solving the optimisation problem, this method allowing the expected value in Equation 4.4 to be not everywhere differentiable. It is beyond the scope of this dissertation to examine further the subgradient method; it is easily implemented and integrated with the general approach proposed in this dissertation, albeit at an increased computational cost.

We can generalise the above approach by removing some restrictions on g. One approach we propose is to suppose there exists $G : \Omega \longrightarrow \mathbb{R}^q$ given by $(x_1, ..., x_q) \mapsto (g_1(x_1), ..., g_q(x_q))$ where each g_i is a non-decreasing real function. A loss function can then be defined as

$$\|G(\boldsymbol{x}) - G(\boldsymbol{y})\|_{p,\beta}.$$

In the case $p = 1$, Theorem 2 implies that the optimal point estimate is equal to that of Equation 4.4. The weakness of this approach is the strict assumption on G; if each g_i is thought of as a utility function then G does not allow the utility of an outcome to depend on any interaction between the $x_i s$. Of course this assumption can be lifted and a more general form of G allowed, but this would in general not yield an analytical form for the optimal point predictor.

Thus, we conclude that, in a multivariate setting, loss functions of the GPL type are of limited practical use, with the exception of a few special cases. This contrasts with the Bregmann loss functions which generalise easily to the multivariate case.

CHAPTER 5

Penalised parameter estimation

Consider again the objective function

$$S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} s(x_t, \boldsymbol{\theta}),$$

which is minimised with respect to $\boldsymbol{\theta}$ to arrive at a parameter estimate $\widehat{\boldsymbol{\theta}}_T$. Under this minimisation approach, each individual score $s(x_t, \widehat{\boldsymbol{\theta}}_T)$ depends on x_t both directly through the first argument and indirectly through $\widehat{\boldsymbol{\theta}}_T$. As a result, the parameter estimate is adjusted to minimise the estimate of the expected score in the case where the value being forecast is actually known in advance. Therefore $S_T(\boldsymbol{x}_{1:T}, \widehat{\boldsymbol{\theta}}_T)$ is typically an over-optimistic estimate of the expected score (Dawid, 1984) and hence $\widehat{\boldsymbol{\theta}}_T$ may be a poor choice for accurate forecasting. Put differently, the approach of minimising the objective function will often result in overfitting. This problem is often solved by the addition of a penalty function to the objective function (Bickel et al., 2006). Intuitively, the penalty function penalises the complexity of the model and thus serves to reduce overfitting. Penalised estimation is commonly applied in regression; see, for example, Tibshirani (1996), but has applications far beyond this; Bickel et al. (2006) provide some examples. There is also the additional advantage of simpler models being more easily

interpreted. This chapter proposes a fairly general approach to penalised estimation for HMM and begins by building upon the idea of consistency given in Chapter 3 to explain the benefits of estimation over a restricted parameter space. This estimation approach is then extended to penalised estimation, and a range of penalties proposed.

5.1 Estimation over a restricted parameter space

The approach developed in this section follows that of Chapter 3, which explains the importance of establishing consistency of the extremum estimator $\widehat{\boldsymbol{\theta}}_T$. This is the property that a value of $\boldsymbol{\theta}$ minimising $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$, that is $\widehat{\boldsymbol{\theta}}_T$, will in some sense tend to a value of $\boldsymbol{\theta}$ minimising $S_0(\boldsymbol{\theta})$. The utility of this property is that the estimator meets the forecaster's want to minimise the expected score of a forecast.

A key assumption required to establish consistency is that $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ should tend almost surely to $S_0(\boldsymbol{\theta})$ as the sample size tends to infinity. More precisely

$$\sup_{\boldsymbol{\theta}\in\Theta}\{|S_T(\boldsymbol{X}_{1:T}, \boldsymbol{\theta}) - S_0(\boldsymbol{\theta})|\} \to 0 \text{ a.s. as } T \to \infty.$$

This assumption is perhaps not surprising; in order for $\widehat{\boldsymbol{\theta}}_T$ to be near a minimum of $S_0(\boldsymbol{\theta})$ it must be that $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ itself is close to $S_0(\boldsymbol{\theta})$, for any $\boldsymbol{x}_{1:T}$. Of course this requirement, and indeed Theorem 3, are purely asymptotic; for some fixed value of T, no probabilistic bound for $\sup_{\boldsymbol{\theta}\in\Theta}\{|S_T(\boldsymbol{X}_{1:T}, \boldsymbol{\theta}) - S_0(\boldsymbol{\theta})|\}$ is implied by the theorem, where by probabilistic bound is meant a probability of the form

$$\Pr\left\{\sup_{\boldsymbol{\theta}\in\Theta} |S_T(\boldsymbol{X}_{1:T}, \boldsymbol{\theta}) - S_0(\boldsymbol{\theta})| > \epsilon\right\},$$

where $\epsilon \in \mathbb{R}_0^+$. Nonetheless, given that the sample size is always finite, such a bound is of great importance for any practical forecasting; it describes probabilistically the 'nearness' of $S_T(\boldsymbol{X}_{1:T}, \boldsymbol{\theta})$ to $S_0(\boldsymbol{\theta})$. The forecaster should

thus be critically concerned with this bound in order to understand and control the estimate of the expected score. In particular, knowledge of the dependence of such a bound on certain controllable variables may allow for the bound to be reduced for each value of ϵ or, equivalently, the rate of uniform convergence increased (Vapnik, 1992). One such controllable variable is the parameter space Θ. Suppose the parameter space is restricted to the subspace $\Theta_s \subseteq \Theta$. Then for any realisation $\boldsymbol{x}_{1:T}$ of $\boldsymbol{X}_{1:T}$,

$$\sup_{\boldsymbol{\theta}\in\Theta_s} |S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta}) - S_0(\boldsymbol{\theta})| \leq \sup_{\boldsymbol{\theta}\in\Theta} |S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta}) - S_0(\boldsymbol{\theta})|. \tag{5.1}$$

It then follows immediately that, for any $\epsilon \in \mathbb{R}_0^+$,

$$\Pr\left\{\sup_{\boldsymbol{\theta}\in\Theta_s} |S_T(\boldsymbol{X}_{1:T}, \boldsymbol{\theta}) - S_0(\boldsymbol{\theta})| > \epsilon\right\} \leq \Pr\left\{\sup_{\boldsymbol{\theta}\in\Theta} |S_T(\boldsymbol{X}_{1:T}, \boldsymbol{\theta}) - S_0(\boldsymbol{\theta})| > \epsilon\right\}.$$

For many choices of Θ_s, the inequality in Equation 5.1 is strict. Thus the forecaster can attempt to decrease the value of this bound by making a restricted choice for Θ_s; 'restricted' here meaning any $\Theta_s \subset \Theta$.

The following question now arises: if the forecaster can make $S_T(\boldsymbol{X}_{1:T}, \boldsymbol{\theta})$ 'nearer' to $S_0(\boldsymbol{\theta})$ by restricting the parameter space, why not impose a severely restricted parameter space? The answer is contained in the following inequality:

$$\min_{\boldsymbol{\theta}\in\Theta_s} S_0(\boldsymbol{\theta}) \geq \min_{\boldsymbol{\theta}\in\Theta} S_0(\boldsymbol{\theta}); \tag{5.2}$$

if Θ_s is too restrictive then it may be that Θ_s no longer contains any points in Θ_0, in which case the inequality in Equation 5.2 is strict. This overall problem is analogous to the bias-variance trade-off encountered in supervised learning; see, for example, Hastie et al. (2009). By restricting the parameter space overfitting may be reduced but at the risk of the HMM being unable to capture fully the important components of the observed series.

Thus a balance must be sought between producing consistent estimates and capturing fully the components of the observed series. In practice, this balance is achieved by fitting the model L times over a number of nested parameter spaces $\Theta_{s_1} \subset \Theta_{s_2} \subset ... \subset \Theta_{s_L} = \Theta$ (Vapnik, 1992). The final model is then chosen by some model selection criterion; for our purposes, that

which minimises an unbiased estimate of the expected score. It is emphasised that the forecaster is not directly concerned with the particular subspace chosen; rather, this choice determines in part the expected score, which is of direct concern to the forecaster.

It is perhaps useful at this point to provide an example of how this approach is commonly applied. For this purpose, we digress slightly by discussing briefly model selection. When fitting a particular HMM by maximisation of the likelihood, it is common to fit the model for a various number of hidden states; see, for example, Zucchini and MacDonald (2009). Suppose a model is fitted for every entry in $(m_{(1)}, ..., m_{(L)})$ which is an ordered vector of natural numbers of length L. Let $\Theta_{(i)}$ be the parameter space associated with the $m_{(i)}$-state model. Then it is clear that

$$\Theta_{(1)} \subset \Theta_{(2)} \subset ... \subset \Theta_{(L)}.$$

For $i \leq j$, the maximum likelihood for the $m_{(j)}$-state model is necessarily greater than or equal to the likelihood for the $m_{(i)}$-state model, but at the cost of an increased number of parameters. A common approach to model selection, for example, Zucchini and MacDonald (2009), is to select the model with the lowest Akaike information criterion (AIC) or Bayesian information criterion (BIC) value. To add clarity we demonstrate this approach using 10000 simulated observations from a three-state Poisson-HMM. HMMs are fitted to the first 150 observations for $m = 2, ..., 6$ using a minus log-likelihood score; for each value of m, we define the 'training score' as the value of $S_T(\boldsymbol{x}_{1:150}, \widehat{\boldsymbol{\theta}}_{150})$. The score of each HMM on the remaining observations, $S_T(\boldsymbol{x}_{151:10000}, \widehat{\boldsymbol{\theta}}_{150})$, is then calculated. We term this the value the 'testing score' and, due to the large size of the dataset, regard it as a good estimate of the actual score of each model; the actual score of the model is equal to $\mathbb{E}[s(X_t, \widehat{\boldsymbol{\theta}}_{150})]$. The resulting training scores, testing scores and AICs of the fitted HMMs (scaled by a factor of 1/300) are compared in Figure 5.1. Overfitting is clearly visible in the divergence between the training and testing scores for $m > 3$. AIC helps to correct the training score; it matches better the shape of the testing score curve.

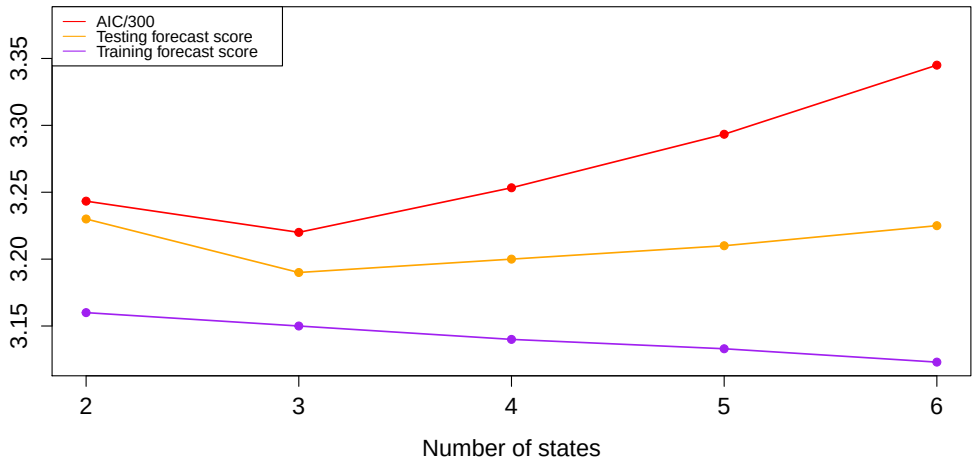


Figure 5.1: Comparison of in-sample and out-of-sample scores for different number of states.

5.2 Extension to penalised extremum estimators

The previous chapter described how restricting the parameter space can improve the estimate of the expected score. Critical to applying this approach is the selection of a suitable restricted subspace. The above approach of fitting the model for various numbers of hidden states is a simple example of how to pick subspaces. We now propose a broader approach which restricts not only the number of hidden states but also the parameters of the state-dependent distributions and the transition probabilities. For example, the forecaster may impose the restriction that no transition probability exceeds 0.5. However, a key difficulty with this approach is that it is not obvious how then to pick useful subspaces of Θ. Thus, we propose that an easy approach to selecting subspaces is to pick a single very simple subspace and then allow parameter values which are sufficiently close to this subspace.

5.2.1 The penalised estimation equation

Required for this purpose is a function which measures the 'nearness' of a particular value of $\boldsymbol{\theta}$ to the simple subspace of Θ. We term this function a 'penalty function' and propose the following definition.

Definition 1. *Let $\Theta_s \subseteq \Theta$. The mapping $J : \Theta \longrightarrow \mathbb{R}_0^+$ is called a **penalty function** if there exists a premetric[1] d on Θ such that $J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}_s \in \Theta_s} d(\boldsymbol{\theta}, \boldsymbol{\theta}_s)$ for all $\boldsymbol{\theta} \in \Theta$.*

This simple definition compresses a number of desirable properties. First, no penalty is incurred if $\boldsymbol{\theta}$ is in the restricted parameter space; $J(\boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \Theta_s$. Second, if $\boldsymbol{\theta}$ is not in the restricted parameter space, the definition ensures that the penalty increases the further $\boldsymbol{\theta}$ is away from the nearest value in Θ_s; 'further' here is meant in terms of the premetric d. The utility of Definition 1 is the simplicity and intuitive appeal of characterising any penalty by just two components: the 'simple' subspace Θ_s, and the measure of discrepancy d.

Suppose now it is required that any estimate of $\boldsymbol{\theta}$ is sufficiently close, in terms of J, to Θ_s. The estimation problem is then

$$\begin{aligned} \underset{\boldsymbol{\theta} \in \Theta}{\text{minimize}} \quad & S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta}) \\ \text{subject to} \quad & J(\boldsymbol{\theta}) \leq r, \end{aligned}$$

where $r \in \mathbb{R}_0^+$; each value of r corresponds to a particular subspace of Θ. For optimisation problems of the above form, the Karush-Kuhn-Tucker (KKT) conditions, (Karush, 1939) and Kuhn and Tucker (1951), are often employed to demonstrate a particular solution is optimal. Problematic in this case is that $S_T(\cdot, \boldsymbol{\theta})$ is not differentiable and thus does not satisfy the standard KKT conditions. Nor is the objective function convex, which excludes extensions of the KKT conditions for convex and non-differentiable functions; see, for example, Ruszczyński (2006).

[1] The mapping $d : \Theta \times \Theta \longrightarrow \mathbb{R}_0^+$ is a premetric if is satisfies $d(x, x) = 0$ and $d(x, y) \geq 0$ for all $x, y \in \Theta$.

Fortunately, analogous conditions to those of KKT for the general form of objective function considered here are provided by Clarke (1976). These conditions give the parameter estimate as

$$\widehat{\boldsymbol{\theta}}_{T,\alpha} \in \underset{\boldsymbol{\theta}\in\Theta}{\arg\min} \left\{ S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta}) + \alpha J(\boldsymbol{\theta}) \right\}, \tag{5.3}$$

where J is a penalty function and $\alpha \geq 0$ a tuning parameter, which adapts the size of the penalty to a specific sequence of observations; methods for determining α are discussed in Section 7.2. Furthermore, for any value of $\alpha > 0$, the corresponding value of r is given by the following identity

$$\alpha(J(\widehat{\boldsymbol{\theta}}_{T,\alpha}) - r) = 0.$$

The utility of this approach is the conversion from a constrained to unconstrained optimisation problem; albeit with the condition that $\boldsymbol{\theta} \in \Theta$. Equation 5.3 also justifies the use of the term 'penalty function' to describe $J(\boldsymbol{\theta})$; the discrepancy of $\boldsymbol{\theta}$ from Θ_s is 'penalised' by increasing the value of the quantity being minimised.

5.3 A selection of penalties for HMMs

This section proposes several forms of the penalty function J; that is, we propose several parameter subspaces Θ_s, as well as a distance measure d, and then find the resulting penalty function $J(\boldsymbol{\theta})$ by solving

$$J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}_s\in\Theta_s} d(\boldsymbol{\theta}, \boldsymbol{\theta}_s).$$

Examples of penalised estimation for HMMs in the literature are rare, and restricted to minus log-likelihood score functions. For example, Städler and Mukherjee (2013) and McGibbon et al. (2014) propose applications of the graphical lasso (Friedman et al., 2008) and fusion L_1 penalty (Tibshirani et al., 2005) to the multivariate Normal-HMM respectively. Keller and Lutz (2002) provide an example of maximum *a posteriori* (MAP) estimation for an HMM; an approach equivalent to penalised likelihood estimation. The key is

that these examples tailor the penalties for a specific application. In contrast, we present a fairly general approach that is not restricted to a particular application. Nonetheless, we emphasise that the purpose of the penalty is to increase the forecast accuracy. Thus, while a penalty with an interpretable form may be more desirable, it is not a requirement. Finally, we note that the standard regression penalties can be easily applied when estimating an HMM with covariates; we do not pursue these particular penalties further.

This section proceeds by considering penalties on only the parameters of the state-dependent distributions, and just the t.p.m. Penalties on the entire parameter vector are then considered, and a proposal made for a Kullback-Leibler based penalty. Finally, a note on notation: if a parameter vector is subscripted by some symbol (e.g. $\boldsymbol{\theta}_s$) then the corresponding t.p.m. is superscripted by the same parendissertationed symbol (e.g. $\boldsymbol{\Gamma}^{(s)}$) and the corresponding matrix of state-dependent distributions' parameters denoted similarly (e.g. $\boldsymbol{\lambda}^{(s)}$).

5.3.1 Penalties on state-dependent distributions

We begin by developing penalties analogous to those in standard regression, the forms of which are

$$\widehat{\boldsymbol{\beta}} = \underset{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{p+1}}{\arg\min} \left\{ \|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}\boldsymbol{\beta}\|_2 + \alpha J(\boldsymbol{\beta}) \right\},$$

where $\boldsymbol{y}_{n\times 1}$ is a response vector, $\boldsymbol{X}_{n\times p}$ a normalised matrix of independent variables and $\boldsymbol{\beta}$ the parameter of interest. An important penalty is the least absolute shrinkage and selection operator (LASSO), Tibshirani (1996), where the penalty is the L_1 norm. In terms of Definition 1, the LASSO penalty follows from setting $\Theta_s = \mathbf{0}$ and d is the metric induced by the L_1 norm. Another important example is ridge regression, Hoerl (1962), where instead d is the semimetric induced by the square of the L_2 norm.

Direct application of these penalties to an HMM is ill-advised as the penalties are designed for tasks which relate a response variable to input variables, whereas HMMs aim to identify structure in the data. Nonetheless,

observe that both penalise deviation from the zero vector; the value of $\boldsymbol{\beta}$ for that model in which the predictors have no influence on the response. This can be thought of as the simplest possible model, analogous to the null hypothesis in hypothesis testing. Thus, one approach to penalised HMMs is to adopt the general principle of this penalty; that is, penalise deviation from the simplest possible model.

Put differently, a natural starting place is to develop a penalty function which penalises discrepancy between a considered HMM and the simplest possible HMM. This requires a strict definition of what is meant by the 'simplest possible HMM'. Thus consider the forecaster's justification for using an HMM, which is: given some time-ordered sample, a single standard distribution is deemed insufficient due to the apparent presence of sub-populations. An independent mixture model (which can be regard as an HMM with every row of the t.p.m. being equal) is then introduced as a result. However, this model fails to account for serial correlations in the data. The rows of the t.p.m. are then allowed to vary, resulting in the HMM. Both of these steps introduce additional complexity in the model which may lead to overfitting and, if so, should be penalised. Hence, the simplest case is proposed to be the assumption that the data are generated from a single instance of the common state-dependent distribution. This is equivalent to assuming all the λ_is are equal. Thus, Θ_s is the subspace of Θ such that $\lambda_i = \lambda_j$ for all $i, j \in M$. It is emphasised that no restrictions are placed on $\boldsymbol{\Gamma}$; transitions between states have no affect on the model if all the λ_is are equal.

Now, letting d be the metric induced by the L_p norm on Θ it is possible to find $J(\boldsymbol{\theta})$. Suppose that for any $\boldsymbol{\theta}_s \in \Theta_s$ the common state-dependent

distribution parameter is denoted bv λ. Then

$$\begin{aligned} J(\boldsymbol{\theta}) &= \min_{\boldsymbol{\theta}_s \in \Theta_s} d(\boldsymbol{\theta}, \boldsymbol{\theta}_s) \\ &= \min_{\boldsymbol{\theta}_s \in \Theta_s} \left\{ \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(s)}\|_p + \|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^{(s)}\|_p \right\} \\ &= \min_{\lambda \in \Lambda} \|\boldsymbol{\lambda} - \mathbf{1}_{m\times 1}\lambda'\|_p + \|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}\|_p \\ &= \|\boldsymbol{\lambda} - \mathbf{1}_{m\times 1}\widehat{\lambda}'\|_p, \end{aligned} \tag{5.4}$$

where $\widehat{\lambda} = \arg\min_{\lambda \in \Lambda} \|\boldsymbol{\lambda} - \mathbf{1}_{m\times 1}\lambda'\|_p$; note that $\widehat{\lambda}$ is d-dimensional. This result is fairly intuitive; deviation from some measure of central tendency, $\widehat{\lambda}$, is penalised. The key is that unlike the LASSO or ridge regression, the measure of central tendency is not fixed at zero but rather adapts itself automatically to the estimate of $\boldsymbol{\lambda}$.

We will concern ourselves with two important cases of the above penalty, namely when $p = 1$ and $p = 2$. In the former case, $\widehat{\lambda}$ is the median of $\boldsymbol{\lambda}$ and, in the later it is the mean of $\boldsymbol{\lambda}$. The differences between these two penalties are explained well by a graph of the constraint regions. Consider first the case where $\boldsymbol{\lambda}_{2\times 1} = (\lambda_1, \lambda_2)$ and suppose it is required that $J(\boldsymbol{\theta}) \leq 0.5$. Setting $\widehat{\lambda} = 0$ yields the well known constraint regions for the LASSO and ridge regression; if $p = 1$ the constraint region is an oblique square centred at the origin, and if $p = 2$ it is disk centred at the origin. These regions are shown in Figure 5.2. If $\widehat{\lambda}$ adapts itself as in Equation 5.4, then the above constraint regions are extended along the line given by $\lambda_1 = \lambda_2$; that is, if $p = 1$ the constraint region is formed by the combination of oblique squares centred at every point such that $\lambda_1 = \lambda_2$. If $p = 2$, the constraint region is formed from circles centred at these same points. The forms of both regions inside the unit square are shown in Figure 5.3.

The similarity between the two shapes disguises the differences present at higher dimensions. To illustrate this, shown in Figure 5.4 are the same constraint regions in the unit cube for $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$. For $p = 1$, the constraint region is an infinite hexagonal prism centred about the line of points satisfying $\lambda_1 = \lambda_2 = \lambda_3$. For $p = 2$, the constraint region is an infinite

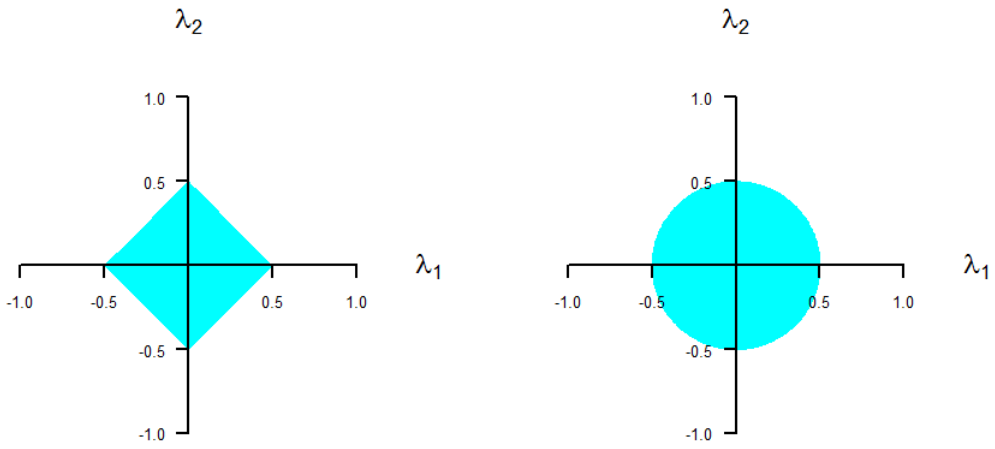


Figure 5.2: Two-dimensional constraint regions for $p = 1$ and $p = 2$ respectively, with $\widehat{\lambda} = 0$.

cylinder centred about the same set of points. Note the the regions in Figure 5.4 have had the corner sections removed to reveal the cross section.

It would be useful at this point to illustrate the behaviour of these two penalties with real data. For this purpose we consider a series of annual counts of major earthquakes between 1900 and 2006 (Zucchini and MacDonald, 2009). For both penalties, we fit Poisson-HMMs for a range of α values; here $p(x|\lambda_i) = \lambda_i^x e^{-\lambda_i}/x!$. The values of the λ_is for different values of α, termed the parameter profile of $\boldsymbol{\lambda}$, are shown in Figure 5.5 for $p = 1$ and in Figure 5.6 for $p = 2$. Note that as α increases both penalties set equal the λ_is and thus the dimension of $\boldsymbol{\lambda}$ is reduced. It is notable that the parameter profiles do not progress smoothly in α; in most cases a pair of parameters is set equal when the distance between them falls below a certain threshold, the resulting being discrete 'jumps' in the parameter profile.

In terms of the parameter space, the L_1 penalty shows a natural progression of $\boldsymbol{\lambda}$ from four dimensions to three then two dimensions and finally a single-state model. In contrast, the L_2 shows an unstable progression of $\boldsymbol{\lambda}$ to-

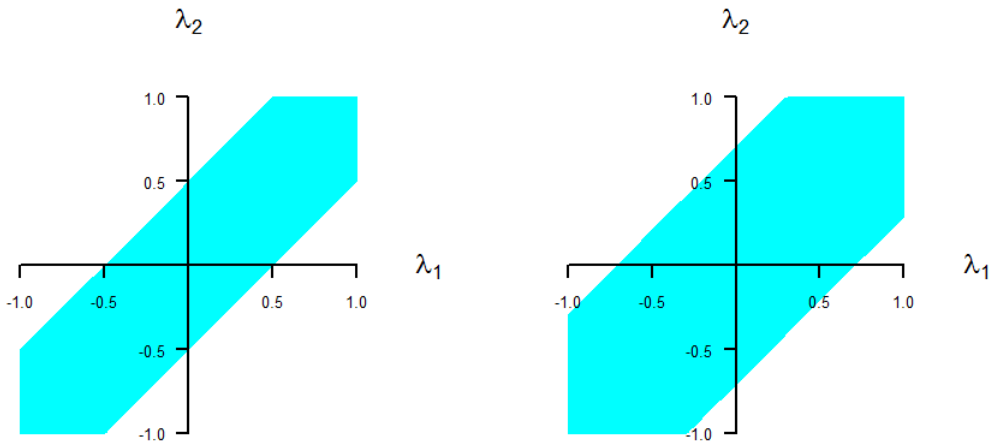


Figure 5.3: Two-dimensional constraint regions for $p = 1$ and $p = 2$ respectively, with adaptive $\widehat{\lambda}$.

wards three dimensions, with oscillations between three and four dimensions. No two dimensional vector is ever attained, three dimensions go immediately into a single-state model.

We provide an explanatory example for this phenomenon. Suppose that instead a three-state model was fitted; the general behaviour of the parameter profiles in these cases can be seen in Figures 5.5 and 5.6 by discarding the region where four distinct λ_i are present. In addition, assume that, without loss of generalisation, $\lambda_1 \leq \lambda_2 \leq \lambda_3$. Observe that both Figures 5.5 and 5.6 show a similar behaviour of λ_1 and λ_3. They key difference is in the behaviour of λ_2; for $p = 1$, λ_2 quickly tends toward λ_1 whereas, for $p = 2$, three distinct values persist until a single-state is reached. Thus, for the purpose of this example, suppose that λ_1 and λ_3 are fixed. We show how the two penalties influence the value of λ_2. The L_1 penalty function is given by

$$|\lambda_1 - \lambda_2| + |\lambda_2 - \lambda_2| + |\lambda_3 - \lambda_2| = \lambda_3 - \lambda_1,$$

which does not depend on λ_2. Put differently, λ_2 can take on any value

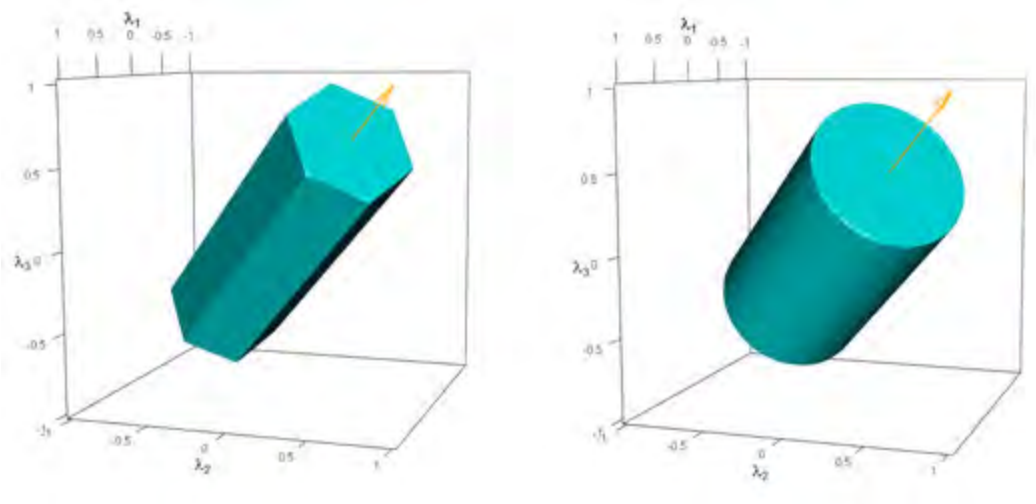


Figure 5.4: Three-dimensional constraint regions for $p = 1$ and $p = 2$ respectively, with adaptive $\widehat{\lambda}$.

in $[\lambda_1, \lambda_3]$ without altering the penalty. On the other hand, the L_2 penalty function is given by

$$\sum_{i=1}^{m} \left(\lambda_i - \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} \right).$$

The above term is convex in λ_2 with a minimum at $\lambda_2 = (\lambda_1 + \lambda_3)/2$; that is, the L_2 penalty actively encourages three distinct values for the λ_i. This is a critical conclusion and, if some form of parameter space reduction is desirable, favours heavily the L_1 penalty.

Finally, we emphasise the difference between distinct values of the λ_is and distinct states; two states i and j cannot be 'merged' if $\lambda_i = \lambda_j$ as the respective transition probabilities may differ.

5.3.2 Penalties on t.p.m.s

The previous section focused on penalties for $\boldsymbol{\lambda}$, here attention is given to penalties for the t.p.m. $\boldsymbol{\Gamma}$. Penalties on $\boldsymbol{\Gamma}$ are particularly important as a m-state HMM requires $dm + m(m-1)$ parameters to be estimated; d is the dimension of each λ_i. Of these parameters, $m(m-1)$ are transition

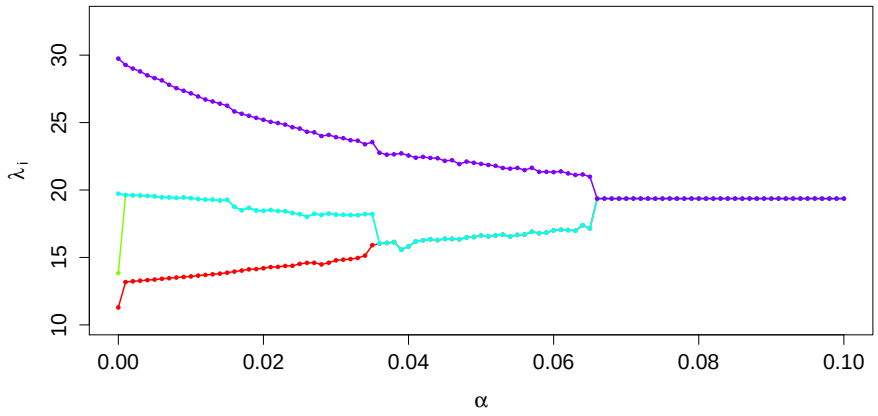


Figure 5.5: Parameter profile of $\boldsymbol{\lambda}$ for L_1 penalty.

probabilities; the majority unless d is large or m is very small.

To begin, we return to restricted parameter spaces which, in this context, are derived by imposing a structure on the t.p.m. By 'structure' is meant any condition on the t.p.m. $\boldsymbol{\Gamma}$ such that the resulting parameter space Θ_s is a subspace of Θ. Structuring of the t.p.m. is present in the HMM literature but is fairly uncommon; see, for example, Cooper and Lipsitch (2004), Langrock et al. (2012) and Bulla and Bulla (2006). The attraction of structuring is the possibility of a large reduction in the number of parameters to be estimated, especially when m is large. There are many possible structures for t.p.m.s, we present three potentially useful examples:

1. the tridiagonal structure $\boldsymbol{\Gamma}_{\text{tri}}$ where non-zero elements are only allowed on the main diagonal, subdiagonal and superdiagonal of the t.p.m. Put differently,
$$\gamma_{ij} = 0 \text{ if } |i-j| > 1 \text{ for } i,j \in M;$$
2. the doubly stochastic structure $\boldsymbol{\Gamma}_{\text{d.s.}}$ where the transpose of t.p.m. is also a t.p.m. Mathematically,
$$\sum_{k=1}^{m} \gamma_{ki} = 1 \text{ for } i \in M;$$

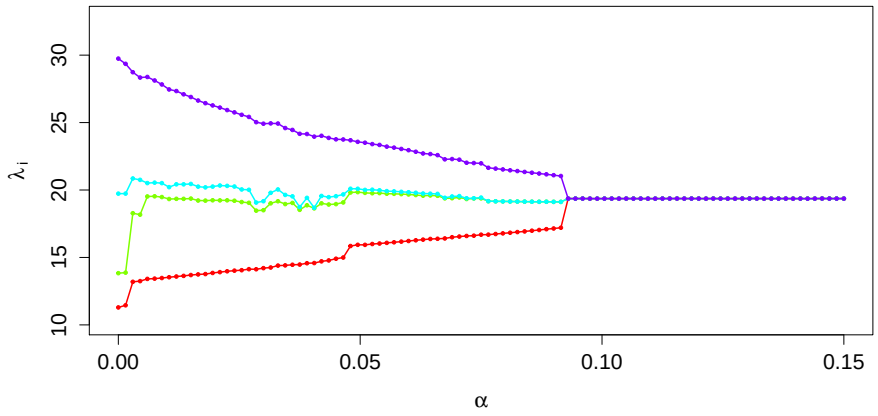


Figure 5.6: Parameter profile of $\boldsymbol{\lambda}$ for L_2 penalty.

3. a single parameter structure $\boldsymbol{\Gamma}_{\text{Dahl}}$ due to Dahl (2004) where

$$\gamma_{ij} = \begin{cases} 1-\gamma & \text{if } i=j=1 \text{ or } i=j=m, \\ 1-2\gamma & \text{if } i=j, i \notin \{1,m\}, \\ \gamma & \text{if } |i-j|=1, \\ 0 & \text{otherwise.} \end{cases}$$

for $\gamma \in [0, 0.5]$.

These conditions are all additional to the standard conditions of a t.p.m: $\gamma_{ij} \geq 0$ and $\sum_{k=1}^{m} \gamma_{ik} = 1$ for $i, j \in M$. The first two structures may hold intuitive appeal but this is certainly not required for a particular structure; at most we require the structure be a useful empirical tool for decreasing the expected score.

We now propose extending structuring to penalised estimation by choice of a suitable penalty function J over the restricted parameter space Θ_s; we term this approach 'soft structuring'. More specifically, we characterise some penalty J by taking Θ_s to be the parameter subspace implied by a particular t.p.m. structure and d as some premetric. Under this approach, the t.p.m. is encouraged, but not forced, to follow a particular structure; hence the name 'soft structuring'.

For the purposes of this dissertation, d is taken as the L_1 premetric. Thus, for any $\boldsymbol{\theta}_s \in \Theta_s$, $d(\boldsymbol{\theta}, \boldsymbol{\theta}_s) = \sum_{i=1}^{m} \sum_{j=1}^{m} |\gamma_{ij} - \gamma_{ij}^{(s)}|$ where $\boldsymbol{\Gamma}^{(s)} = (\gamma_{ij}^{(s)})$ is the t.p.m. associated with $\boldsymbol{\theta}_s$. Given below is the form of J for this choice of d and the three structures above; proofs are given in Lemmas 4 to 6 in Appendix B.3.

1. For the tridiagonal structure the penalty is given by

$$J(\boldsymbol{\theta}) = 2 \sum_{i,j \in M : |i-j|>1} \gamma_{ij};$$

that is, the sum of elements not on the main, sub or super-diagonal.

2. For the doubly-stochastic structure the penalty is given by

$$J(\boldsymbol{\theta}) = \sum_{j=1}^{m} \|\boldsymbol{\Gamma}_{\bullet j} - \mathbf{1}\|_1 = \sum_{j=1}^{m} \left| \sum_{i=1}^{m} \gamma_{ij} - 1 \right|;$$

that is, the sum of the absolute differences between each column total and one.

3. For the Dahl structure the penalty is given by

$$\begin{aligned} J(\boldsymbol{\theta}) = {} & |\gamma_{11} + \widehat{\gamma} - 1| + |\gamma_{mm} + \widehat{\gamma} - 1| + \sum_{i=2}^{m-1} |\gamma_{ii} + 2\widehat{\gamma} - 1| \\ & + \sum_{i,j \in M : |i-j|=1} |\gamma_{ij} - \widehat{\gamma}| + \sum_{i,j \in M : |i-j|>1} \gamma_{ij}, \end{aligned} \tag{5.5}$$

where $\widehat{\gamma}$ is the 'estimated' value of γ given by

$$\widehat{\gamma} = \min\{\widehat{\boldsymbol{\gamma}}_{1/2}, 0.5\}.$$

Here $\widehat{\boldsymbol{\gamma}}_{1/2}$ denotes the median of the values γ implied by each element on the main, sub and superdiagonal of $\boldsymbol{\Gamma}$.

5.3.3 Penalties based upon the Kullback-Leibler divergence

The previous two subsections proposed penalties for $\boldsymbol{\lambda}$ and $\boldsymbol{\Gamma}$ separately; no attention was given to penalties for the entire parameter vector. Nonetheless,

such penalties should be considered as penalisation was proposed as a method of improving the forecast accuracy of the HMM. Thus, while separately both the $\boldsymbol{\lambda}$ and $\boldsymbol{\Gamma}$ penalties have intuitive appeal, it is not required that a penalty should be intuitive. Hence, a penalty on the entire parameter vector must be considered as it may provide a better forecast accuracy than the previously considered penalties. In principle, any two penalties on $\boldsymbol{\lambda}$ and $\boldsymbol{\Gamma}$ may be combined provided the same measure of distance is used. However, when considered separately, the scaling of $\boldsymbol{\lambda}$ and $\boldsymbol{\Gamma}$ is not of major concern. The difficulty when combining penalties is that the scaling becomes relevant. For example, the order magnitude of the λ_is may well be significantly higher than the γ_{ij}s.

This difficulty arises due to a more general problem associated with using the the L_1 or L_2 metric for penalties. That problem is, though the L_p metrics are simple and correspond well to known forms of regularisation, for example the ridge and LASSO penalty, they are arguably a poor measure of discrepancy between two HMMs. The reason for this is that these metrics do not take into account the probabilistic behaviour of the model; only the distance between the parameters is measured.

An alternative approach is to set d as a suitable measure of discrepancy between two hidden Markov models. By 'suitable' is meant a measure that examines the difference in the probability measure of two HMMs resulting from differences in the parameter vector, as opposed to simply examining the difference between the parameter vectors. For this purpose we propose use of a statistical 'divergence' which measures the distance between two probability distributions. The key is that a divergence is a not a direct measure of distance between parameters, but rather a measure of distance between the probability measures implied by the value of the parameters.

In particular, we consider the well-studied Kullback-Leibler divergence (KLD), due to Kullback and Leibler (1951). For two HMMs parameterised by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and on a continuous sample space, the KLD of $\boldsymbol{\theta}_2$ from $\boldsymbol{\theta}_1$ is

given by

$$D_{\mathrm{KL}}(\boldsymbol{\theta}_1 \| \boldsymbol{\theta}_2) = \int_{\Omega} \mathbb{P}_{\boldsymbol{\theta}_1}(x) \log \frac{\mathbb{P}_{\boldsymbol{\theta}_1}(x)}{\mathbb{P}_{\boldsymbol{\theta}_2}(x)} \, \mathrm{d}x. \tag{5.6}$$

This function is asymmetric, the common interpretation being that the first argument corresponds to the 'true' model and the second to an approximation of the first. For our purposes, we will take $\boldsymbol{\theta}_1$ as the parameter vector corresponding to an unrestricted model, and $\boldsymbol{\theta}_2$ as the parameter vector corresponding to a restricted model; that is, a model with a restricted parameter space. In this sense, the KLD can be interpreted as a measure of how well the simple model approximates the more complex one. Note that, for a discrete sample space, the integral is simply converted to a sum.

Equation 5.6 does not, in general, have a closed-form solution; rather a solution must be found numerically. This approach, however, is unsuitable if the integral is to be evaluated many times; for example, during an optimisation routine where $D_{\mathrm{KL}}(\boldsymbol{\theta}_1 \| \boldsymbol{\theta}_2)$ is included in the penalty function. Instead, we follow the approach of Ling and Dai (2012) by approximating the the divergence using an upper bound. For the HMM, Do (2003) gives an upper bound of the form

$$D_{\mathrm{KL}}(\boldsymbol{\theta}_1 \| \boldsymbol{\theta}_2) \leq \sum_{i=1}^{m} \delta_i^{(1)} \left(D_{\mathrm{KL}} \left(\lambda_i^{(1)} \| \lambda_i^{(2)} \right) + D_{\mathrm{KL}} \left(\boldsymbol{\Gamma}_{i\bullet}^{(1)} \| \boldsymbol{\Gamma}_{i\bullet}^{(2)} \right) \right), \tag{5.7}$$

where $D_{\mathrm{KL}} \left(\lambda_i^{(1)} \| \lambda_i^{(2)} \right)$ is the KLD between $p(x|\lambda_i^{(1)})$ and $p(x|\lambda_i^{(2)})$, and

$$D_{\mathrm{KL}} \left(\boldsymbol{\Gamma}_{i\bullet}^{(1)} \| \boldsymbol{\Gamma}_{i\bullet}^{(2)} \right) = \sum_{j=1}^{m} \gamma_{ij}^{(1)} \log \frac{\gamma_{ij}^{(1)}}{\gamma_{ij}^{(2)}}. \tag{5.8}$$

Equation 5.7 has a closed-form solution if the KLD between the state-dependent distributions also has a closed-form solution. This is true for many families of distributions; for example, the Poisson family, the Gamma family, the normal family and binomial family when the number of trials is fixed (Burnham and Anderson, 2002). This upper bound is also fairly intuitive and shows a clear separation between the divergences arising from differences in the state-dependent distributions and in the t.p.m; nonetheless, there is an interaction due to $\boldsymbol{\delta}$ being a function of $\boldsymbol{\Gamma}$.

To form a penalty function, the KLD upper bound is combined with various restricted parameter-spaces. Suppose again that Θ_s is the subspace of Θ such that $\lambda_i = \lambda_j$ for all $i, j \in M$. Then

$$J(\boldsymbol{\theta}) = \min_{\lambda \in \Lambda} \sum_{i=1}^{m} \delta_i D_{\mathrm{KL}}(\lambda_i \| \lambda).$$

This optimisation problem often has a closed-form solution. For example, in the Poisson-HMM case, it is easily shown that the value of λ minimising the above term is

$$\lambda = \sum_{i=1}^{m} \delta_i \lambda_i,$$

and thus the final penalty is given by

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{m} \delta_i D_{\mathrm{KL}} \left(\lambda_i \| \sum_{i=1}^{m} \delta_i \lambda_i \right). \tag{5.9}$$

The KLD may also be used for soft structuring of the t.p.m. A caveat in this case is that the KLD requires $\gamma_{ij}^{(1)} = 0$ if and only if $\gamma_{ij}^{(2)} = 0$; see Equation 5.8. Thus certain structures may not be used if distance measure in the penalty is taken as the KLD; for example, the tridiagonal structure. Of course, an advantage of using the KLD for penalties on the entire parameter-vector is that the problem of scaling discussed in Section 5.3.3 falls away.

CHAPTER 6

Cross-validation for hidden Markov models

In Chapter 5 we introduced the penalty function $\alpha J(\boldsymbol{\theta})$, which consists of two components: the penalty function J and the tuning parameter α. The interpretation of α was discussed in Section 5.2 but little has been said about how α should be calculated. This chapter seeks to remedy that omission by describing a cross-validation approach for this purpose. First, a general introduction to cross-validation is given. Then a number of cross-validation schemes are presented, following which a simulation study is described. Finally, some alternatives to cross-validation are described.

6.1 An introduction to cross-validation

Consider again the objective of the forecaster; that is, to find an estimate of $\widehat{\boldsymbol{\theta}}_{T,\alpha}$ that minimises the expected score,

$$S_0(\widehat{\boldsymbol{\theta}}_{T,\alpha}) = \mathbb{E}[s(X_t, \boldsymbol{\theta}_{T,\alpha})] = \int_\Omega s(x, \widehat{\boldsymbol{\theta}}_{T,\alpha}) \, \mathrm{d}F(x),$$

where F is the the actual cumulative distribution function of x. To approximate this expected value, the objective function $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ was proposed.

However, if $\widehat{\boldsymbol{\theta}}_{T,0}$ is a value of $\boldsymbol{\theta}$ minimising this objective function, then

$$\mathbb{E}[S_T(\boldsymbol{X}_{1:T}, \widehat{\boldsymbol{\theta}}_{T,0})] \leq S_0(\widehat{\boldsymbol{\theta}}_{T,0});$$

that is, $S_T(\boldsymbol{x}_{1:T}, \widehat{\boldsymbol{\theta}}_{T,0})$ is a biased estimate of $S_0(\widehat{\boldsymbol{\theta}}_{T,0})$. This is because the expected score is being estimated using a value of $\boldsymbol{\theta}$ defined to minimise the objective function which makes it likely that overfitting will occur. For this reason, $S_T(\boldsymbol{x}_{1:T}, \widehat{\boldsymbol{\theta}}_{T,\alpha})$ is termed the in-sample score; indicating that the score is calculated on the same sample used to estimate the parameters. Similarly, if $\boldsymbol{x}_{t,t+k}$ is another set of observations distinct from $\boldsymbol{x}_{1:T}$ then $S_T(\boldsymbol{x}_{t,t+k}, \widehat{\boldsymbol{\theta}}_{T,\alpha})$ is termed the out-of-sample score as the observations in $\boldsymbol{x}_{t,t+k}$ were not used to fit the model. Finally, we term the expected score $S_0(\widehat{\boldsymbol{\theta}}_{T,\alpha})$ the 'actual score'.

The bias of the in-sample score suggests that a method that provides an estimated parameter closer to Θ_0 is required. For this purpose a penalty function was proposed, which aims to ensure that $S_0(\widehat{\boldsymbol{\theta}}_{T,\alpha}) \leq S_0(\widehat{\boldsymbol{\theta}}_{T,0})$. Thus, to determine α a better estimate of the actual score than $S_T(\boldsymbol{x}_{1:T}, \widehat{\boldsymbol{\theta}}_{T,\alpha})$ is needed. We will use a cross-validation approach for this purpose.

Heuristically, cross-validation involves dividing the observed data into a training set and a validation set. The HMM is fitted on observations in the training set and the out-of-sample score calculated for the observations in the validation set. This process may be repeated for different training and validation sets, and the cross-validation score calculated as the weighted average of the out-of-sample score on each validation set.

6.1.1 A general framework for cross-validation

We begin by describing more formally the process for selecting a value of α. Thus, let $G(1), ..., G(K)$ be K mutually exclusive subsets of the index set $\{1, ..., T\}$, with $T_i = |G(i)|$ for $i \in \{1, ..., K\}$. Denote by $\boldsymbol{x}_{-G(i)}$ the data $\boldsymbol{x}_{1:T}$ with the jth observation regarded as missing if $j \in G(i)$, similarly $\boldsymbol{x}_{G(i)}$ denotes the data with all but the jth observation regarded as missing, where $j \in G(i)$. For example, suppose that five observations are

made and that $\boldsymbol{x}_{1:5} = (4, 8, 1, 3, 7)$. Letting $G(1) = \{2, 3\}$ implies that $\boldsymbol{x}_{-G(1)} = (4, \text{NA}, \text{NA}, 3, 7)$ and $\boldsymbol{x}_{G(1)} = (\text{NA}, 8, 1, \text{NA}, \text{NA})$; NA denotes a missing observation.

The vectors $\boldsymbol{x}_{-G(i)}$ and $\boldsymbol{x}_{G(i)}$ are the training and validation sets respectively. The penalised parameter estimate for the case when the observations in $G(i)$ are regarded as missing is given by

$$\widehat{\boldsymbol{\theta}}_{-G(i),\alpha} \in \underset{\boldsymbol{\theta}\in\Theta}{\arg\min} \left\{ S_T(\boldsymbol{x}_{-G(i)}, \boldsymbol{\theta}) + \alpha J(\boldsymbol{\theta}) \right\}. \tag{6.1}$$

As $\widehat{\boldsymbol{\theta}}_{-G(i),\alpha}$ is found without knowledge of the observations in $\boldsymbol{x}_{G(i)}$, the out-of-sample score calculated on these observations can be regarded as an estimate of the actual score. If there are multiple training sets, then a weighted average of the out-of-sample scores can be calculated. More precisely, let S_{T_i} denote the usual extremum estimator objective function for T_i observations, then an estimate of the actual score is

$$\text{CV}(\alpha) = \frac{1}{\sum_{j=1}^{K} T_j} \sum_{i=1}^{K} T_i S_{T_i}(\boldsymbol{x}_{G(i)}, \widehat{\boldsymbol{\theta}}_{-G(i),\alpha}), \tag{6.2}$$

which is termed the cross-validation score. As stated above, α should be selected to minimise the value of Equation 6.2. Thus consider the vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_L)$; a finite set of possible values for α. The chosen value of α is given by

$$\widehat{\alpha} = \underset{\alpha\in\boldsymbol{\alpha}}{\arg\min}\, \text{CV}(\alpha). \tag{6.3}$$

Noting that each value of α corresponds to some restricted parameter space, this approach follows that of Vapnik (1992) where the model is fitted over a number of restricted subspaces, and the final model chosen to minimise an estimate of the actual score. As an aside, we prefer to avoid calling $\widehat{\alpha}$ the 'optimal' value of α as consideration is given to only a finite set of α values. It should be emphasised that the purpose of cross-validation is to determine the value of $\widehat{\alpha}$ only. Once a value has been chosen, the final model is fitted using all the data with $\alpha = \widehat{\alpha}$.

Irrespective of the selection of the $G(i)$s, there are four desirable properties for a particular cross-validation scheme in the time-series context. The

first and second are adequacy and diversity of the training and validation sets respectively (Tashman, 2000). By adequacy is meant the property of having a sufficient number of observations for estimating accurately the parameters and the out-of-sample score. By diversity is meant the property of the estimated parameters and out-of-sample score not depending on some special event in the observations. A particular set is adequate and diverse provided there are a high number of observations in the set and these observations are spread across the observation period. The third desirable property is zero correlation between the training and validation sets. If the validation set depends strongly on the training set, then the resulting score cannot be regarded as out-of-sample. The fourth desirable property is that the time-ordering of the observations is not disrupted by the cross-validation scheme.

6.2 Some cross-validation schemes

The usefulness of stating cross-validation in terms of Equation 6.3 is that emphasis is given to the fact that the only decision is the selection of the form of the $G(i)$s. This section proceeds by discussing a number of different ways for making this choice. First, cross-validation for i.i.d. data is described and its shortfalls in a time-series context discussed. Consideration is then given to two methods for general time-series cross-validation and two methods specific to hidden Markov models.

6.2.1 Basic cross-validation for i.i.d. data

Before describing cross-validation schemes for time-series, it will be helpful to provide a brief introduction to cross-validation for i.i.d. data and highlight the shortfalls of this approach in a time-series context. The standard approach for i.i.d. data is to partition randomly $\{1, ..., T\}$ into K subsets $G_1, ..., G_K$ of roughly equal size; see, for example, Hastie et al. (2009). Thus the model is fitted K times with different training and validation sets each time. This ensures that all the data are used to fit and test the model which helps

achieve adequacy and diversity. There are, however, two problems with using this approach for cross-validation of an HMM. First, the time-order of the observations is disrupted by the random partitioning. Second, the validation set may not be independent of the training set as the observations are not independent.

6.2.2 Last-block validation

To avoid the problems with basic cross-validation, Hjorth (1993) proposes the use of last-block validation. That is, the first t observations are taken as the training set and the remaining observations as the validation set. In the notation of Equation 6.1, $K = 1$ and $G(1) = \{t + 1, ..., T\}$; for brevity we will just write G as opposed to $G(1)$. Thus the model is fitted and the score measured on the validation set only once. Bergmeir and Benítez (2012) suggest selecting t in such a way that approximately 20% of the observations are in the validation set.

There are a number of advantages of this approach. First, it represents accurately how the forecaster will apply the model in practice. Second, the time-order of the observations is not disrupted. Finally, while there will be some dependence between the initial observations in $\boldsymbol{x}_G$ and the training set, the effect should be small for a sufficiently large validation set. The problem with this approach is the possibility of the training set not being adequate and diverse. Inadequate because only a portion of the data is used to calculate the out-of-sample score, and not diverse as there may be special events that occur only towards the end of the observation period.

It should be emphasised that the parameters are not recalibrated for each observation in G; the model is fitted only once. There are two reasons for this. First, it is computationally infeasible to refit the model for each observation in G, especially when the number of states is large. Second, not recalibrating the model is arguably more relevant for applications as models are often built and calibrated once over a particular period of forecasting (Bergmeir and Benítez, 2012).

6.2.3 Cross-validation with hv-blocks

As a solution to the problems present in both basic cross-validation and last-block validation, Racine (2000) proposed 'hv-block cross-validation' which modifies basic cross-validation to help ensure that the time order is not disrupted and the validation set is independent of the training set. This is done by dividing the observations into ordered 'blocks' and then removing observations to decrease the correlation between observations in the training and validation sets. The approach presented here is a modification on that of Racine (2000), whose approach is computationally infeasible for HMMs as it requires the model to be fitted the same number of times as there are observations.

More precisely, divide the index set $\{1, ..., T\}$ into ordered blocks of roughly size v; for example, $G(1) = \{1, 2, ..., v\}$ and $G(2) = \{v + 1, v + 2, ..., 2v\}$. For each time the model is fitted on $\boldsymbol{x}_{-G(i)}$, regard as missing from $\boldsymbol{x}_{-G(i)}$ the h observations on either side of $\boldsymbol{x}_{G(i)}$. For example, if $v = 5$ and $h = 1$, then x_6 and x_7 should be regarded as missing when validating the model on $G(2)$. The value of v controls the size of each validation set; a value which results in five validation sets is considered standard Bergmeir and Benítez (2012). If $v = 1$, Burman et al. (1994) propose setting h as some fixed fraction of the sample size; the rule-of-thumb given is $h/T = 0.25$. Racine (2000) suggest that this rule-of-thumb extends to the more general hv-block case.

The advantages of this approach are that most of the data are used for training and validation, the time-order is maintained and near-independence of the training and validation sets is enforced. The disadvantages are the decrease in training set size when h is large, and the computational burden of fitting the model K times.

The restricted case where $h = 0$ is termed v-block validation, and the case where $v = 1$ is termed h-block validation; see Burman (1989) and Burman et al. (1994) respectively. The weakness of the former is that the dependence between the training and validation sets is not mitigated. The weakness of

the latter is the high computational cost.

6.2.4 Half-sampling and Δ-sequential sampling

Celeux and Durand (2008) propose an HMM-specific cross-validation scheme termed 'odd-even half-sampling' (OEHS). The observations are divided into two sets, an odd set where observations time-indexed by even numbers are regarded as missing, and an even set where observations time-indexed by odd numbers are regarded as missing. The model is then fitted to these two sets, and the out-of-sample score calculated on the alternative set. The final cross-validation score is taken as the weighted average of the two resulting out-of-sample scores. In the notation of Equation 6.1, $G(1) = \{1, 3, ..., 2\left\lceil \frac{T}{2} \right\rceil - 1\}$ and $G(2) = \{2, 4, ..., 2\left\lfloor \frac{T}{2} \right\rfloor\}$; those are, the odd and even sub-sequences respectively.

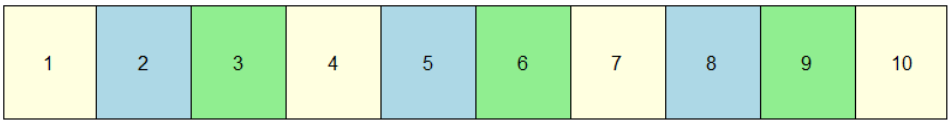
The key advantage of this approach is that diversity is aided by spreading all the training and validation sets across all the observations. The computational requirements are also low as the model is fitted only twice. A disadvantage is that the dependence between the training and validation, which could be significant, cannot be removed. In addition, the training sets may be small if there are few observations.

This last weakness is significant; the training sets contain only half the observations, a small proportion in comparison to the other cross-validation schemes. As a solution, we propose a new cross-validation scheme which generalises OEHS, which we term 'Δ-sequential sampling' (DSS) where Δ is some natural number greater than two. Under a DSS scheme, there are Δ validation sets spread over the observation period where the index for each validation set forms a sequence with a constant difference of Δ; the ith index set begins at i. If $\Delta = 2$, then DSS is equivalent to OEHS. DSS can thus be regarded as a generalisation of OEHS, allowing for more than two validation sets. Alternatively, it may be viewed as an alteration of v-block validation in which each validation set is spread over most of the observation period.

More formally, for $i \in \{1, 2, .., \Delta\}$, the validation sets are given by

$$G(i) = \{i, i + \Delta, i + 2\Delta, ..., \Delta \left\lfloor \frac{(T - i)}{\Delta} \right\rfloor + i\}.$$

For example, if $T = 10$ and $\Delta = 3$, then validation sets will look as follows, where each colour indicates a validation set.



The key advantage of DSS over OEHS is the increase in the size of each training set, which is approximately $\frac{(\Delta-1)T}{\Delta}$. A choice of Δ equal to three, four or five results in training set sizes in-line with the previously suggested cross-validation schemes.

6.3 A correction term for the cross-validation score

To conclude this section, we adopt an idea proposed by Burman et al. (1994) in the context of h-block validation. Observe first that, provided $\boldsymbol{x}_{G(i)}$ is independent of $\boldsymbol{x}_{-G(i)}$ for every i,

$$\mathbb{E}[\mathrm{CV}(\alpha)] = \frac{1}{K} \sum_{i=1}^{K} \int_{\Omega} s(x, \widehat{\boldsymbol{\theta}}_{-G(i),\alpha}) \, \mathrm{d}F(x).$$

The cross-validation term can then be expanded as follows

$$\begin{aligned}\mathrm{CV}(\alpha) =& \frac{1}{\sum_{j=1}^{K} T_j} \sum_{i=1}^{K} T_i S_{T_i}(\boldsymbol{x}_{G(i)}, \widehat{\boldsymbol{\theta}}_{-G(i),\alpha}) - \mathbb{E}[\mathrm{CV}(\alpha)] \\ &+ \mathbb{E}[\mathrm{CV}(\alpha)] - S_0(\widehat{\boldsymbol{\theta}}_{T,\alpha}) + S_0(\widehat{\boldsymbol{\theta}}_{T,\alpha}) \\ =& \left(\frac{1}{\sum_{j=1}^{K} T_j} \sum_{i=1}^{K} T_i S_{T_i}(\boldsymbol{x}_{G(i)}, \widehat{\boldsymbol{\theta}}_{-G(i),\alpha}) - \frac{1}{K} \sum_{i=1}^{K} \int_{\Omega} s(x, \widehat{\boldsymbol{\theta}}_{-G(i),\alpha}) \, \mathrm{d}F(x) \right) \\ &+ \int_{\Omega} \left(\frac{1}{K} \sum_{i=1}^{K} s(x, \widehat{\boldsymbol{\theta}}_{-G(i),\alpha}) - s(x, \widehat{\boldsymbol{\theta}}_{T,\alpha}) \right) \mathrm{d}F(x) + S_0(\widehat{\boldsymbol{\theta}}_{T,\alpha}).\end{aligned}$$

As the forecaster wants $\mathrm{CV}(\alpha)$ to be as close as possible to $S_0(\widehat{\boldsymbol{\theta}}_{T,\alpha})$, Burman et al. propose adding a correction term to the cross-validation score to bring it closer to $S_0(\widehat{\boldsymbol{\theta}}_{T,\alpha})$.

The first term in brackets cannot be removed as the approximation for the integral component is simply the cross-validation score. In addition, if $\boldsymbol{x}_{G(i)}$ is independent of $\boldsymbol{x}_{-G(i)}$ for every i, then the expectation of this term is zero. However, the second term can, at best, be approximated using the available observations as

$$\frac{1}{\sum_{j=1}^{K} T_j} \sum_{i=1}^{K} T_i S_T(\boldsymbol{x}_{1:T}, \widehat{\boldsymbol{\theta}}_{-G(i),\alpha}) - S_T(\boldsymbol{x}_{1:T}, \widehat{\boldsymbol{\theta}}_{T,\alpha}). \tag{6.4}$$

The negation of this term can be added to the cross-validation term for any scheme with the aim of improving the approximation to the out-of-sample score. Thus, unless stated otherwise, by 'cross-validation score' is meant the corrected cross-validation score which is given by

$$\mathrm{CCV}(\alpha) = \mathrm{CV}(\alpha) + S_T(\boldsymbol{x}_{1:T}, \widehat{\boldsymbol{\theta}}_{T,\alpha}) - \frac{1}{\sum_{j=1}^{K} T_j} \sum_{i=1}^{K} T_i S_T(\boldsymbol{x}_{1:T}, \widehat{\boldsymbol{\theta}}_{-G(i),\alpha}).$$

6.4 A simulation study

A short simulation study is now presented with the purpose of illustrating the proposed cross-validation schemes. The simulation set-up was as follows. The

parameters for 5000 m-state stationary Poisson-HMMs were randomly generated [1]. For each HMM, a sequence of 20000 observations was generated. The first 150 observations were used to calculate the cross-validation score using five methods: last-block validation (`LB`), v-blocked validation with $v = 30$ (`VB`), hv-blocked validation with $v = 30$ and $h = 10$ (`HVB`), half-sampling cross-validation (`OEHS`), and finally Δ-sequential sampling with $\Delta = 3$ (`DSS`). The 151st to 250th observations were discarded, and the remaining 19750 observations used to calculate the out-of-sample score for the estimated model; that is,

$$S_T(\boldsymbol{x}_{251:20000}, \widehat{\boldsymbol{\theta}}_{T,\alpha}).$$

Due to the large size of this testing set, the resulting out-of-sample scores may be regard as good estimates of the actual scores. The minus log-likelihood loss function is used to measure the score and, as the simulations are purely illustrative, only unpenalised models are fitted.

To analyse the simulations, the cross-validation score and out-of-sample scores are compared. The resulting difference arises from two discrepancies. First is the difference caused by the estimated parameters on the training sets, the $\widehat{\boldsymbol{\theta}}_{-G(i),\alpha}$, differing from the estimated parameter on all the observations, $\widehat{\boldsymbol{\theta}}_{T,\alpha}$, the effect of which tends to decrease as the training set size increases. Second is the difference between the out-of-sample scores on the validation sets, $S_T(\boldsymbol{x}_{G(i)}, \widehat{\boldsymbol{\theta}}_{-G(i),\alpha})$, and the actual scores $S_0(\widehat{\boldsymbol{\theta}}_{-G(i),\alpha})$, the effect of which tends to decrease as the validation set size increases. As data are limited, the effect of these two factors cannot be considered independently of each other. By comparing the cross-validation score with the actual score, these two components are analysed simultaneously; an effective cross-validation scheme should find a balance between these two factors. 'Effective' here

[1] The vector $\boldsymbol{\lambda}$ was generated by randomly selecting uniformly each $\lambda_i$on the interval $(0, 40)$ and, then multiplying λ_i by (i/m). The t.p.m. $\boldsymbol{\Gamma}$ was generated by randomly selecting uniformly each γ_{ij} on the interval $(0, 1)$. The diagonal elements were then multiplied by three, and each row of the matrix scaled to sum to one. The purpose of the modifications to $\boldsymbol{\lambda}$ and $\boldsymbol{\Gamma}$ is to ensure that 'trivial cases' were avoided; for example, when the HMM degenerates to a independent mixture model.

means an accurate estimate of the actual score of the HMM with parameter $\widehat{\boldsymbol{\theta}}_{T,\alpha}$.

The simulation study is performed twice for $m = 2$ and $m = 3$, with $\alpha = 0$. The point plots comparing the out-of-sample score and the cross-validation score for each simulated model are given in Figure 6.1. For both studies, all the cross-validation schemes appear to perform fairly well and show a general trend about the identity line; the set of points where the cross-validation score is equal to the out-of-sample score.

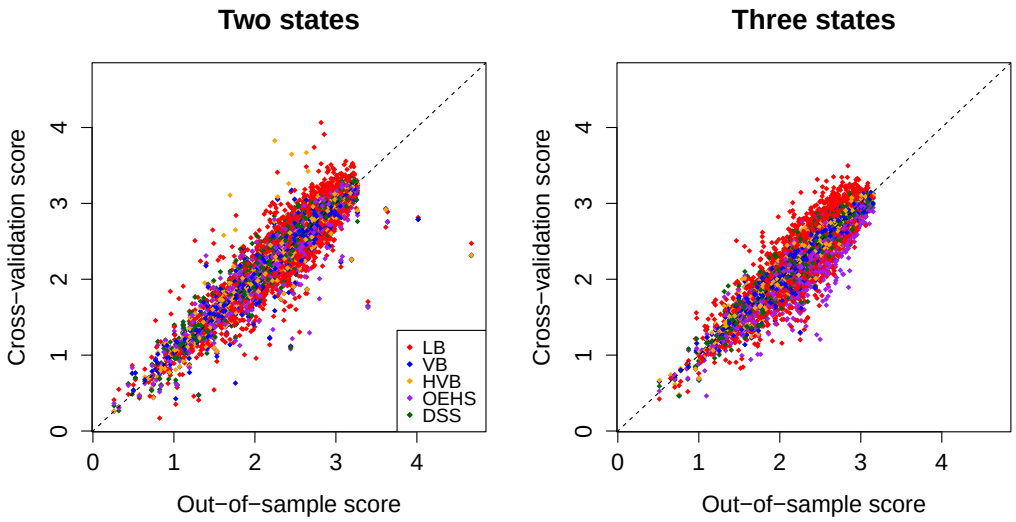


Figure 6.1: Scatter plots of out-of-sample and cross-validation scores for two simulation studies.

To provide further analysis, the cross-validation scores were scaled by dividing each score by the corresponding out-of-sample score. The resulting values are more interpretable in that a value greater than one implies the cross-validation scheme is over-estimating the out-of-sample score, and vice versa for a value less than one. The box plots and summary statistics for the scaled cross-validation scores are given in Figure 6.2 and Table 6.1 respectively [2]. We make the following observations:

[2]The number of simulations, 5000, was determined by assessing when the convergence

- all the schemes perform well, with average scaled cross-validation scores close to one;
- `LB` performs the best in terms of average scaled score but displays a larger variance than the other schemes, an undesirable characteristic;
- `VB` and `HVB` perform well in terms of average scaled score, benefiting from the lowest variance, but incur the largest time cost;
- For $m = 2$, `OHES` performs similarly to `HVB`, but at a greatly reduced time cost. However, the performance for $m = 3$ is poor;
- `DSS` performs relatively poorly for $m = 2$; better for $m = 3$, especially given the time cost compared to `VB` and `HVB`.

	Two states			Three states		
	Average	Variance	Time(s)	Average	Variance	Time(s)
`LB`	0.9999	0.0090	0.0898	1.0002	0.0079	1.6565
`VB`	0.9971	0.0028	0.5454	0.9973	0.0019	10.2908
`HVB`	1.0014	0.0032	0.5450	1.0032	0.0019	10.2275
`OEHS`	1.0014	0.0034	0.2145	0.9791	0.0034	4.3054
`DSS`	1.0084	0.0032	0.3281	0.9981	0.0023	6.4533

Table 6.1: Summary statistics of scaled cross-validation scores for different schemes across two simulation studies.

Finally, it is of interest to determine whether the cross-validation schemes do give significantly different results. Thus, to test formally the differences between the cross-validation schemes, the pairwise differences in the average scaled score within each study were compared. The Wilcoxon signed-rank test[3] (Wilcoxon, 1945) was used to test the hypothesis that the difference

plots of these statistics were consistently flat.

[3]The scaled scores were judged symmetric about the mean by examination of histograms and the box plots.

in average scaled score between two schemes is zero. For each of the two studies, that is for $m = 2$ and $m = 3$, there are ten pairwise differences and thus a total of 20 p-values were calculated. The Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) was then used to determine which null hypotheses to reject; the false discovery rate was set at 0.05. The results of these tests were that all pairwise differences were significantly different from zero, with the exception of the difference between `LB` and `VB` for both $m = 2$ and $m = 3$.

Next, to test the effectiveness of the correction term in Equation 6.4, the cross-validation scores without the inclusion of the correction term are considered; comparative summary statistics are given in Table 6.2. The correction term generally improves the average scaled score, with the exception of `VB` for $m = 2$, and `VB` and `OEHS` for $m = 3$. As above, for each cross-validation scheme in each study, the Wilcoxon signed-rank test (Wilcoxon, 1945) was used to test the hypothesis that the average scaled score with the correction term is equal to the average scaled score without the correction term. The Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) was repeated; the false discovery rate set again at 0.05. Rejected null hypotheses are indicated by boldface of the average scaled score pair in Table 6.2; in this case, all ten of the null hypotheses were rejected, indicating a significant decrease brought by the correction term.

Finally, we analyse how the simulation results change with the inclusion of a penalty function. For this purpose, both studies were repeated with a penalty function of the form

$$\alpha J(\boldsymbol{\theta}) = \alpha \|\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}\|_2,$$

which a heuristic choice made of $\alpha = 0.01$. The resulting summary statistics are given in Table 6.3. For the two-state models, the unpenalised cross-validation schemes tend to outperform the penalised ones. Whereas for the three-state models, the unpenalised cross-validation schemes tend to under perform the penalised ones. This may be explained by considering the ratio of the number of observations to the number of parameters, which is higher

Two states				
	Average scaled score		Variance of scaled score	
	w/ correction	w/o correction	w/ correction	w/o correction
LB	**0.9999**	**1.0020**	0.0090	0.0094
VB	**0.9971**	**0.9996**	0.0028	0.0033
HVB	**1.0014**	**1.0068**	0.0032	0.0117
OEHS	**1.0014**	**1.0211**	0.0034	0.0039
DSS	**1.0084**	**1.0128**	0.0032	0.0031
Three states				
	Average scaled score		Variance of scaled score	
	w/ correction	w/o correction	w/ correction	w/o correction
LB	**1.0002**	**1.0042**	0.0079	0.0082
VB	**0.9973**	**1.0014**	0.0018	0.0019
HVB	**1.0032**	**1.0103**	0.0019	0.0021
OEHS	**0.9791**	**1.0173**	0.0034	0.0037
DSS	**0.9981**	**1.0137**	0.0023	0.0021

Table 6.2: Comparative statistics of scaled cross-validation scores with and without a correction term.

for the two-state models. Thus, for these models, the effect of overfitting is likely to be smaller. To test the differences in average scaled score, the hypothesis testing procedure used for Table 6.2 is repeated in the same exact manner. For both studies the three schemes with the smallest training set sizes, HVB, OEHS and DSS, showed significant differences.

6.5 Alternatives to cross-validation

We conclude this section by discussing two alternatives to the cross-validation approach for selecting α. These alternatives are presented mainly for completeness; cross-validation is computationally feasible for the applications

Two states				
	Average scaled score		Variance of scaled score	
	$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0$	$\alpha = 0.01$
LB	0.9999	1.0004	0.0079	0.0126
VB	0.9971	0.9985	0.0018	0.0032
HVB	**1.0014**	**1.0037**	0.0019	0.0035
OEHS	**1.0014**	**1.0034**	0.0034	0.0037
DSS	**1.0084**	**1.0106**	0.0023	0.0037
Three states				
	Average scaled score		Variance of scaled score	
	$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0$	$\alpha = 0.01$
LB	1.0002	1.0006	0.0079	0.0083
VB	0.9973	0.9976	0.0018	0.0019
HVB	**1.0032**	**1.0022**	0.0019	0.0019
OEHS	**0.9791**	**0.9823**	0.0034	0.0033
DSS	**0.9981**	**1.0000**	0.0023	0.0022

Table 6.3: Summary statistics of scaled cross-validation scores for unpenalised and penalised two- and three-state models.

considered in this dissertation, especially last-block validation and odd-even half-sampling.

The first common alternative is a bootstrap technique; see, for example, the .632+ bootstrap method of Efron and Tibshirani (1997). In an HMM context, as explained in Section 3.4, a parametric, as opposed to non-parametric, bootstrap is required. However, samples generated by a parametric bootstrap are unsuitable for testing out-of-sample performance as the samples are generated from the model being tested.

The second alternative is a theoretically-guided choice, which tends to be problem specific. One example is due to Städler and Mukherjee (2013), who apply the graphical lasso to a multivariate-normal-HMM. Städler and

Mukherjee acknowledge that the computational requirements of cross-validation is high when the number of observations, T, is large and the dimension of the observations q is high; they present an example where $T = 32791$ and $q = 53$. As an alternative, they propose the use of a universal regularisation parameter

$$\alpha_{\text{uni}} = \sqrt{2T \log q}/2,$$

which is a function of T and q only. This form follows from the theoretical work of Friedman et al. (2008), and is specific to the graphical lasso. Thus, it may not be adapted for the general penalties proposed in this dissertation.

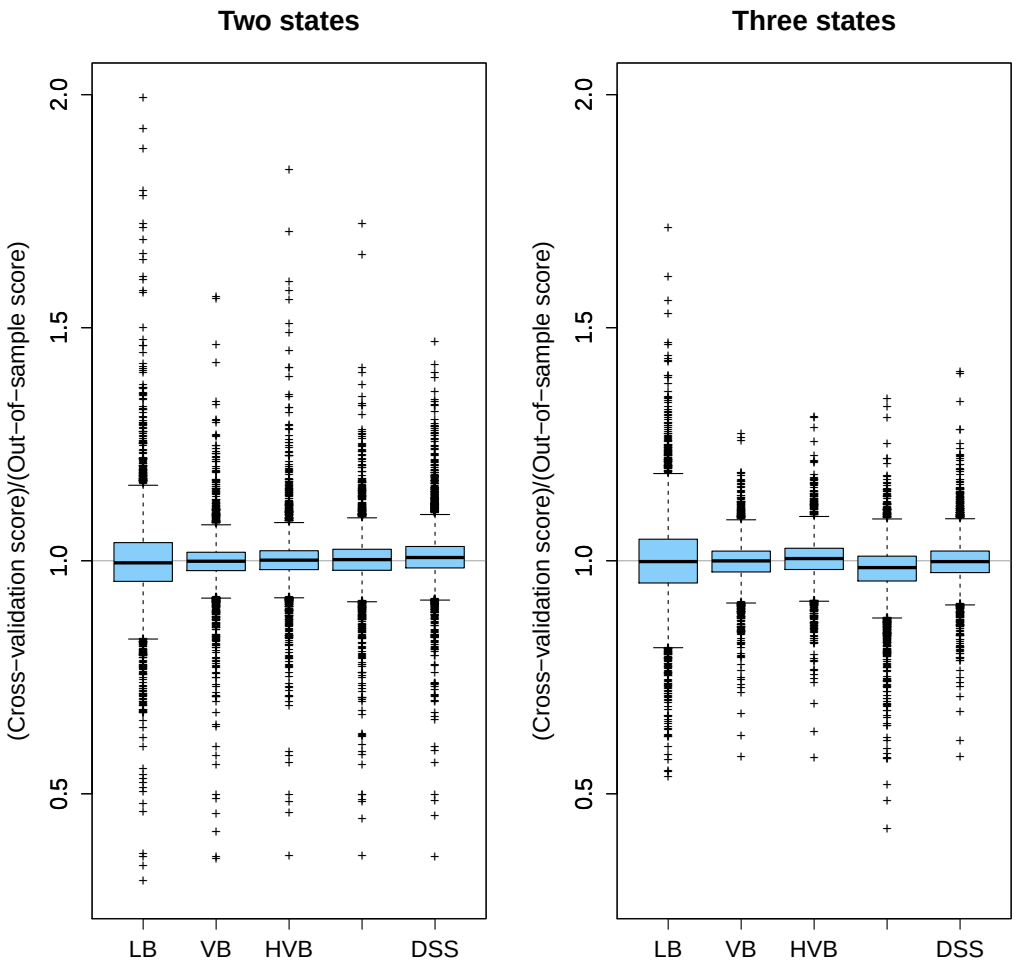


Figure 6.2: Box plots of scaled cross-validation scores for different schemes across two simulation studies.

CHAPTER 7

Model fitting and implementation

We discuss here the fitting of HMMs and implementing the penalties and cross-validation techniques described in Chapters 5 and 6. The general optimisation problem is to find a parameter-vector estimate $\widehat{\boldsymbol{\theta}}_{T,\alpha}$ minimising the objective function

$$f(\boldsymbol{\theta}, \alpha) = S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta}) + \alpha J(\boldsymbol{\theta}). \tag{7.1}$$

We separate this optimisation problem into two components. Considered first is minimising Equation 7.1 via direct numerical minimisation (DNM), for fixed values of α and the number of hidden states. Second, a value of α and the number of hidden states must be chosen. The chapter proceeds by discussing DNM and some computational difficulties that may be encountered. Following this, methods for selecting α and the number of states are described. The chapter concludes by proposing a general approach to forecasting with HMMs.

All calculations performed here, as well as elsewhere in this dissertation, are done so in `R` (R Development Core Team, 2012).

7.1 Direct numerical minimisation of the objective function

Minimisation of Equation 7.1 is, in principle, no different from the standard approach of minimising the unpenalised minus log-likelihood; nuances will arise due to differentiability of $f(\boldsymbol{\theta}, \cdot)$. Thus, for fixed α and m, we minimise the above objective function directly via a numerical minimiser; this is the approach taken by Zucchini and MacDonald (2009) in the unpenalised minus log-likelihood case.

A common alternative to DNM is the EM-algorithm (Dempster et al., 1977), see, for example, Rabiner (1989). As the EM-algorithm is specific to maximum-likelihood based estimation, we would require a more general algorithm; see, for example, the MM-algorithm (De Leeuw and Heiser, 1977). Nonetheless, we do not pursue these algorithms as DNM is deemed sufficient; a discussion of DNM versus EM is given by MacDonald (2014).

We focus first on a number of algorithms which, when given initial parameter values, will iteratively update the current parameter estimate with the aim of reaching a local minimum. Following this, we discuss a heuristic technique which uses these algorithms with the aim of identifying a global minimum.

7.1.1 Statistical packages for DNM

If both $S_T(\cdot, \boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ are everywhere differentiable then a Newton-type algorithm may be used to find a local minimum. One `R` package for this purpose is the unconstrained minimiser `nlm`, which requires transformation of $\boldsymbol{\theta}$; see Zucchini and MacDonald (2009) for details. Alternatively, constrained optimisation is performed by `constrOptim` using one of the following methods: BFGS [1], CF (Fletcher and Reeves, 1964) and L-BFGS-B (Byrd et al., 1995).

[1]The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is named after the four authors who simultaneously published it ; see Broyden (1970), Fletcher (1970), Goldfarb

These approaches may be used when, for example, a minus log-likelihood score is combined with an L_2 penalty on $\boldsymbol{\lambda}$.

Of course, $S_T(\cdot, \boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ are not everywhere differentiable for some of the forms discussed in this dissertation. In these cases a method which does not require derivatives is needed and, for this purpose, we will use the Nelder-Mead method (Nelder and Mead, 1965) which is a fairly simple simplex-based direct search method. A problem with the Nelder-Mead approach is that it tends to converge slowly or may not converge at all (Pham, 2012). We propose a possible solution to this problem. Suppose first that $f(\cdot, \alpha)$ is differentiable almost everywhere. Then `nlm` may provide a good solution, but will tend to fail when a stationary point is at or near a point of non-differentiability. Thus, we propose combining `nlm` with Nelder-Mead. More specifically, `nlm` should be run first to arrive at a solution close to a stationary point, and this solution should then become the starting value for the Nelder-Mead procedure.

7.1.2 Initial values and multiple local minima

A problem with the above techniques is that the objective function will frequently have multiple local minima; we cannot determine if a minimum found by either `nlm` or `constrOptim` is a global minimum. To help resolve this problem, we adopt the approach of Zucchini and MacDonald (2009) whereby the minimisation procedure is performed multiple times with different initial values. The justification for this technique is simple and as follows. The above minimisers are deterministic in the sense that an initial value will *ceteris paribus* lead to the same estimate each time the same minimiser is run. Thus, for each local minimum and minimiser we may associate a set of initial values which, when run through that particular minimiser, produce that particular minimum. A simple example of such sets of initial values is given in Figure 7.1. Is is shown that the parameter space can be divided into two mutually exclusive regions R_1 and R_2 such that points in the same

(1970) and Shanno (1970).

region, if taken as initial values in a optimisation algorithm, lead to the same minimum value.

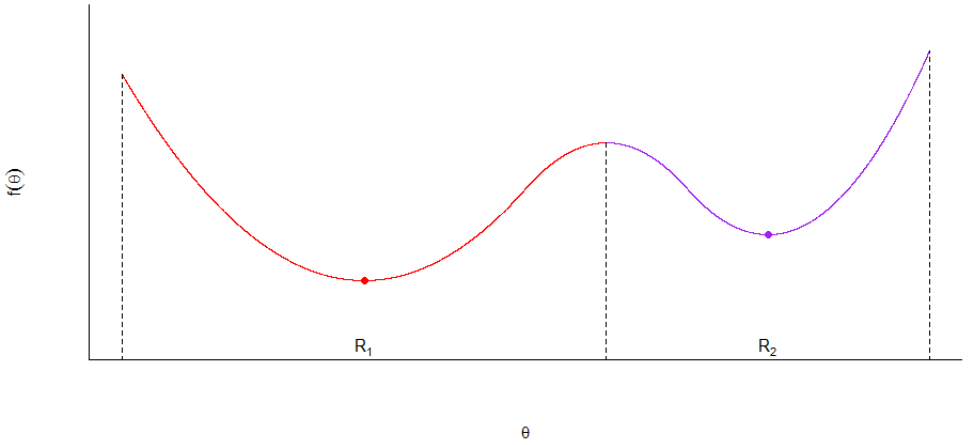


Figure 7.1: Regions of initial values which lead to particular minimum values.

Clearly, if we were able to run a minimiser multiple times, using at least one vector of initial values from each of the aforementioned sets, we will have found the global minimum. Of course, these sets are unknown and at best we can use a wide range of initial values with the hope that at least one results in a global minimum. More precisely, the model is fitted first using a plausible choice of initial values; see Zucchini and MacDonald for details. We term this choice the 'base' initial values, and denote them by $\boldsymbol{\theta}^{(0)}$. All subsequent initial values are then chosen using a proposal distribution conditional on $\boldsymbol{\theta}^{(0)}$; that is,

$$\boldsymbol{\theta}^{(h)} \sim Q(\boldsymbol{\theta}^{(0)}),$$

where Q is the proposal distribution. This sampling process may be repeated a fixed number of times, H, or until no improvement in the minimum achieved is observed for a number of consecutive fits; the total number of fits will tend to increase with the number of states.

The key is that the main purpose of this sampling process is to encourage diversity in the initial values about the first 'plausible choice'. The proposal

distribution Q should be chosen to meet this criterion; for the purposes of this dissertation a normal distribution is used with suitable transformations on the parameters. For example, suppose we fit an m-state Poisson-HMM. Denote the base initial values as $\boldsymbol{\lambda}^{(0)} = (\lambda_i^{(0)})$ and $\boldsymbol{\Gamma}^{(0)} = (\gamma_{ij}^{(0)})$. One sampling scheme may be

$$\log(\lambda_i^{(h)}) \sim \mathcal{N}\left(\log(\lambda_i^{(0)}), |\tfrac{1}{3}\log(\lambda_i^{(0)})|\right) \text{ for } i \in M,$$
$$\log\left(\frac{\gamma_{ij}^{(h)}}{1-\gamma_{ij}^{(h)}}\right) \sim \mathcal{N}\left(\log\left(\frac{\gamma_{ij}^{(0)}}{1-\gamma_{ij}^{(0)}}\right), 3\right) \text{ for } i, j \in M, i \neq j.$$

The choice of variance parameters is very subjective; larger variances will encourage the sampling procedure to explore more of the parameter space.

7.2 Picking a suitable value for the tuning parameter

Having discussed minimising 7.1 for fixed α, we now consider a method for selecting a suitable value of α. We described in Chapter 6 the general approach for selecting α. That is, find, for a range of α values, the cross-validation score of the model. The final value of α chosen should be that which minimises the cross-validation score. This section describes how the range of α values should be chosen and some computational difficulties that may arise.

Selecting a good value of α is analogous, and equivalent in the minus log-likelihood score case, to hyper-parameter optimisation in Bayesian statistics (Bergstra and Bengio, 2012). For problems of that form, a common technique is to combine a manual and grid search; see, for example, Larochelle et al. (2007). More precisely, the forecaster will specify a particular finite range for α, and the model fitted for a grid of α values in that range. This range may be refined with the purpose of improving the estimate of α. Despite its simplicity, grid search is fairly reliable in one-dimension (Bergstra and Bengio, 2012).

In the context of this dissertation, models are fitted over L ordered values

of α; $\alpha_1, ..., \alpha_L$. As above, α_1 and α_L must be specified by the forecaster and the remaining α_is are then set at equally spaced intervals in the range $[0, \alpha_L]$. The smallest value, α_1, is always taken as zero. Unfortunately, selection of the largest value, α_L, is more difficult as it is not clear a priori at which point increasing the size of the penalty function will decrease the forecast accuracy of the model. We propose setting α_L such that penalty function does not exceed half the score of the unpenalised model. That is,

$$\alpha_L = \frac{S_T(\boldsymbol{x}_{1:T}, \widehat{\boldsymbol{\theta}}_{T,0})}{2J(\widehat{\boldsymbol{\theta}}_{T,0})}. \tag{7.2}$$

This rule is very *ad hoc* but worked well for some of the applications considered in Chapter 8. Of course, the range $[0, \alpha_L]$ may be refined as the model is fitted for the various α_is.

The choice of L will depend on the computational resources available; a large value of L may be time consuming. As a method of accelerating this process, we propose a warm-start algorithm that finds parameter estimates sequentially for the entire $\boldsymbol{\alpha}$ vector. The basic idea is that $\widehat{\boldsymbol{\theta}}_{T,\alpha_i}$ will be close to $\widehat{\boldsymbol{\theta}}_{T,\alpha_{i-1}}$ and, thus, having found the latter, we can use it to find the former more quickly. More precisely, the model is first fitted for $\alpha = \alpha_1 = 0$ using a plausible choice of base initial values, that is $\boldsymbol{\theta}^{(0)}$, and the multiple start procedure described in Section 7.1.2. Then, for fitting the model with subsequent α_is, the base initial values are taken as the estimate found when fitting the model with $\alpha = \alpha_{i-1}$; for example, when fitting the model with $\alpha = \alpha_2$ we take $\boldsymbol{\theta}^{(0)} = \widehat{\boldsymbol{\theta}}_{T,0}$. The key is that multiple starts are still made; the problem of multiple local minima is present. However, the optimisation process can be made faster by considering initial values in a region where we expect to find the optimum value.

We describe this algorithm below in full detail. The integer H denotes the number of initial values sampled. The output of algorithm is the set of parameter estimates $\{\widehat{\boldsymbol{\theta}}_{T,\alpha}\}_{\alpha \in \boldsymbol{\alpha}}$.

Algorithm 1: Heuristic optimisation with warm starts.

Initialise $\boldsymbol{\theta}^{(0)}$.

for $i := 1$ **to** L **step** 1 **do**

fit HMM with initial values $\boldsymbol{\theta}^{(0)}$ to obtain $\widehat{\boldsymbol{\theta}}_{T,\alpha_i}$.

for $j := 1$ **to** $H - 1$ **step** 1 **do**

sample initial values $\boldsymbol{\theta}^{(j)} \sim Q(\boldsymbol{\theta}^{(0)})$

fit HMM with initial values $\boldsymbol{\theta}^{(j)}$ to obtain $\widehat{\boldsymbol{\theta}}_{T,\alpha_i,j}$.

if $f(\widehat{\boldsymbol{\theta}}_{T,\alpha_i,j}, \alpha_i) < f(\widehat{\boldsymbol{\theta}}_{T,\alpha_i}, \alpha_i)$ **then set** $\widehat{\boldsymbol{\theta}}_{T,\alpha_i} = \widehat{\boldsymbol{\theta}}_{T,\alpha_i,j}$.

end for

set $\boldsymbol{\theta}^{(0)} = \widehat{\boldsymbol{\theta}}_{T,\alpha_i}$.

end for

7.3 Picking the number of states

As with α, the number of states, m, should be chosen to minimise the cross-validation score. However, unlike α, it is fairly easy to select candidate values of m. The model selection process is as follows. First, the forecaster selects a range of consecutive m values to evaluate. For each m value, Algorithm 1 in Section 7.2 is run; the range of α values will tend to be different for each value of m. The output of this process is a total of Lm parameter estimates. The final estimate is then chosen as the one which minimises the cross-validation score.

7.4 A general approach to forecasting with HMMs

We conclude this chapter by stating a general four-step approach to forecasting with HMMs. This approach follows directly from the content of Chapters 2 to 8, and is as follows. First, the forecaster should determine a score function which accurately describes the loss incurred. Second, a penalty function should be chosen with the aim of reducing the out-of-sample score. Third, the number of states and tuning parameter value should be chosen to minimise the cross-validation score; the HMMs may be fitted using Algorithm 1. The final step is to check the HMM using pseudo-residuals and the potential risk measured.

CHAPTER 8

Applications

In this chapter we present four applications of penalised HMMs. There are three key purposes for these applications. The first is to investigate if matching the measure of forecast accuracy with the method of parameter estimation results in improved forecast accuracy; that is, if extremum estimators are useful for HMMs. The second is to examine if the introduction of a penalty improves the actual score of the model. The third purpose of these applications is to determine if we can use cross-validation to pick an α value which results in a lower actual score.

The following applications include two simulation studies and two applications to real data. Simulation studies are useful as they allow for the actual score to be calculated to a high degree of accuracy and real data applications provide a realistic demonstration of our proposed approach. In both cases, the data are divided into a training set and a testing set. The training set may be regarded as the data available to the forecaster and the testing set will be used to estimate the actual score of the various HMMs. As the approach developed in this thesis is fairly general, we consider a range of different types of HMMs, score functions and penalty functions.

8.1 Simulation study: a univariate categorical-HMM

We present here a simulation study with the purpose of demonstrating the basic approach to forecasting proposed in this dissertation. We consider a series of 20150 observations simulated from a three-state categorical-HMM with five categories; that is, for $i \in \{1, 2, 3\}$

$$p(x|\boldsymbol{\lambda}_i) = (\boldsymbol{\lambda}_i)_x,$$

where $x \in \{1, 2, 3, 4, 5\}$ and $\boldsymbol{\lambda}_i$ is a probability distribution of length five. The parameter matrix $\boldsymbol{\lambda}$ was chosen to be such that the three states corresponded to a 'low', 'medium' and 'high' state. The precise parameter values are as follows:

$$\boldsymbol{\lambda} = \begin{pmatrix} 0.77 & 0.20 & 0.03 & 0 & 0 \\ 0.17 & 0.26 & 0.26 & 0.19 & 0.12 \\ 0.01 & 0.06 & 0.15 & 0.30 & 0.48 \end{pmatrix}; \; \boldsymbol{\Gamma} = \begin{pmatrix} 0.85 & 0.10 & 0.05 \\ 0.10 & 0.80 & 0.10 \\ 0.05 & 0.25 & 0.70 \end{pmatrix}.$$

The first 150 observations simulated are taken as the training set, and the remaining 20000 as the testing set. The total number of parameters to be estimated is 18, of which 12 are state-dependent probabilities. This is a large number of parameters to estimate given only 150 observations, and would probably lead to overfitting in the absence of a penalty.

For this simulation a fairly simple approach is taken; the score function is the minus log-likelihood and we apply an L_2 penalty on $\boldsymbol{\lambda}$. These choices allow the use of `nlm` for minimising the objective functions and aim to demonstrate the usefulness of penalties for a fairly standard application of the HMM. We will consider fitting only a three-state HMM given that we have knowledge of the actual HMM used to generate the data.

We fit a three-state HMM using Algorithm 1; the number of α values considered is 10 and the largest value in $\boldsymbol{\alpha}$, α_{10}, is calculated using Equation 7.2. Ten starting values are used for each minimisation as this number of

starts resulted in a high degree of stability of the parameter estimates. The cross-validation scheme used is Δ-sequential sampling with $\Delta = 3$. The resulting out-of-sample and cross-validation scores are shown in Figure 8.1. We make the following two observations. First, the shape of the out-of-sample

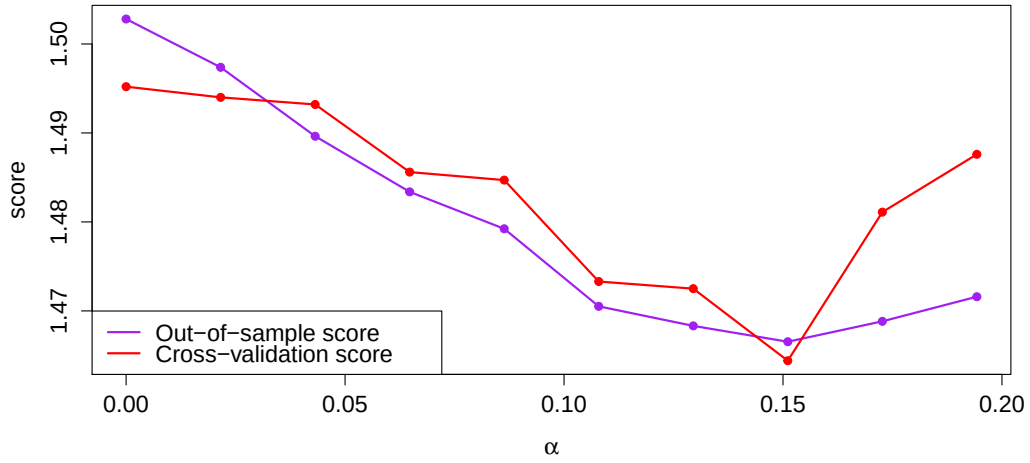


Figure 8.1: Comparison of categorical-HMM's out-of-sample and cross-validation scores for increasing α.

score curve is fairly intuitive; the score decreases initially as the penalty helps reduce overfitting and then starts to increase as penalty becomes too restrictive. Second, the cross-validation score provides a fairly good match to the out-of-sample score, except when α is quite large. Critically, both the cross-validation score and out-of-sample score are minimised for the same value of α, which is 0.151.

Next we examine the usefulness of penalties more closely. For this purpose we shall consider three HMMs: the unpenalised HMM, a penalised HMM with $\alpha = 0.151$ which we term the 'penalised HMM', and the true HMM used to generate the data. The decrease in the out-of-sample score brought by the penalised HMM may appear fairly modest. However one should be wary of interpreting a score function on an absolute basis. In fact, the actual

out-of-sample score of the true HMM is 1.42. Therefore the deviances of the unpenalised and penalised HMMs on the testing set are 24.47 and 13.59 respectively; this suggests a large improvement brought by the penalty.

Next we check the fit of these HMMs using pseudo-residuals. Q-Q plots of the normal randomised pseudo-residuals are shown in Figure 8.2. All three models show a fairly good fit; the Q-Q plot for the penalised HMM appears marginally closer than the plot for the unpenalised does.

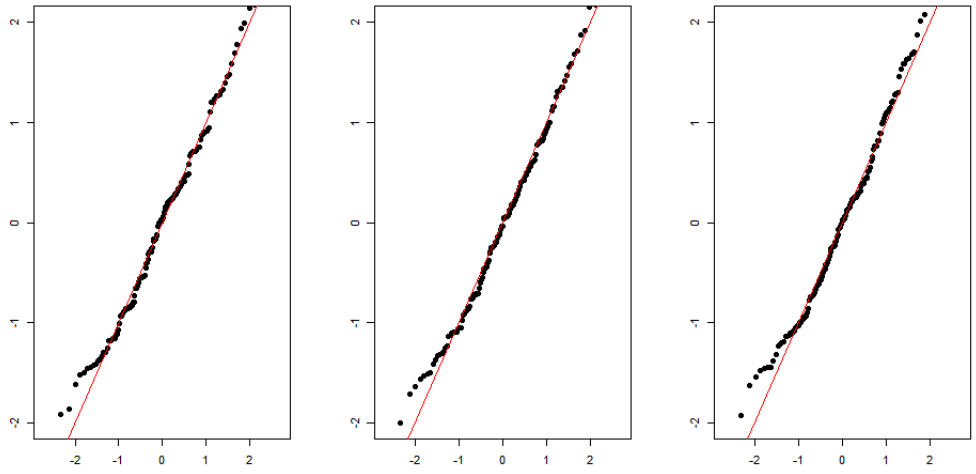


Figure 8.2: Quantile-quantile plots for the unpenalised, penalised and true categorical-HMMs respectively. Theoretical quantiles are shown on the horizontal axis.

Finally, potential risk of both the unpenalised and penalised models is calculated. That is, for each time point t in the training set we calculate the interval described in Equations 3.3 and 3.4 for both $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{T,0}$ and $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{T,0.151}$; a 95% confidence level is used. The plot of these intervals is given in Figure 8.3. The intervals for the penalised model tend be narrower and their midpoints lower; this is preferred by the forecaster. We re-emphasise that the potential risk is not a measure of goodness-of-fit; it describes a probabilistic bound on the score by assuming that the estimated parameters are the 'true' parameters. However we can compare the potential risk for the unpenalised,

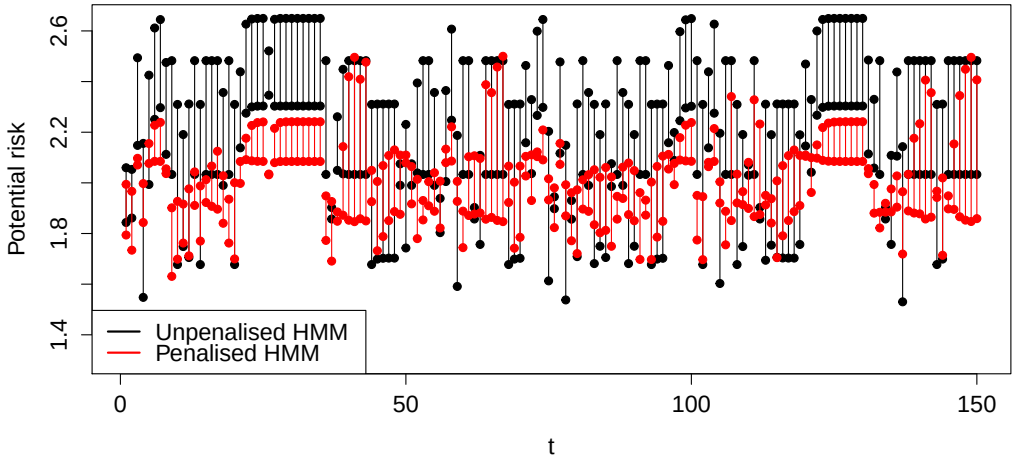


Figure 8.3: Comparison of potential risk intervals for unpenalised and penalised categorical-HMM.

penalised and true models. For ease of interpretation only the midpoints of the intervals are given; see Figure 8.4. Note that the potential risk for the penalised HMM generally fits better the potential risk for the true HMM than does the unpenalised HMM. The exception to this is when the potential risk for the true HMM is very high; this occurs when the Markov chain is in state 1 and the underestimation of the potential risk is probably caused by the penalised HMM overestimating the probability of remaining in state 1.

This simulation study shows how the introduction of a penalty improves the forecast accuracy of the categorical-HMM. Compared to the unpenalised HMM, the penalised HMM shows a decrease in the out-of-sample score, appears to fit better the training data, and results in lower potential risk. In addition, the cross-validation scheme chosen was able to identify a good choice of α.

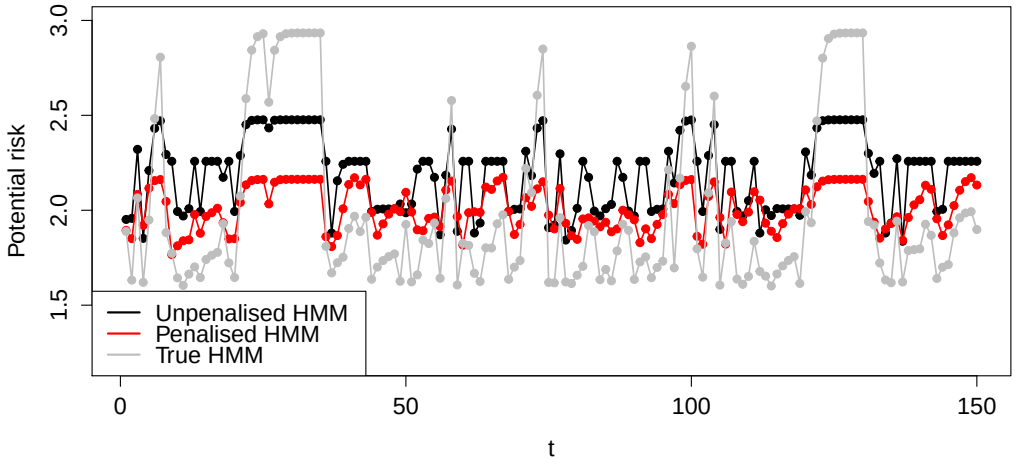


Figure 8.4: Comparison of potential risk midpoints for unpenalised, penalised and true HMM.

8.2 Simulation study: a multivariate exponential-HMM

The focus of this simulation study is on the t.p.m. penalties introduced in Section 5.3.2. The study aims first, to investigate the usefulness of t.p.m. penalties in improving the forecast accuracy of the HMM and, second, to compare penalised estimation with the simpler approach of structuring the t.p.m; that is, imposing a particular structure as opposed to penalising deviation from that structure.

We simulate from a four-state bivariate exponential-HMM for this study. When defining the state-dependent distributions, we assume that the state-dependent joint density is a product of the corresponding marginal probabilities; Zucchini and MacDonald (2009) term this property 'contemporaneous conditional independence'. Therefore, for an observation $\boldsymbol{x}_t = (x_{t1}, x_{t2})$, the joint-density function conditional on the Markov chain being in state i is

given by

$$p_i(\boldsymbol{x}_t) = \lambda_{i1}\lambda_{i2}e^{-\lambda_{i1}x_{t1}-\lambda_{i2}x_{t2}}.$$

The parameter values are as follows:

$$\boldsymbol{\lambda} = \begin{pmatrix} 0.96 & 0.81 \\ 0.30 & 1.32 \\ 0.92 & 0.45 \\ 1.51 & 1.08 \end{pmatrix}; \; \boldsymbol{\Gamma} = \begin{pmatrix} 0.87 & 0.07 & 0.03 & 0.03 \\ 0.07 & 0.75 & 0.08 & 0.10 \\ 0.06 & 0.05 & 0.82 & 0.07 \\ 0.03 & 0.04 & 0.02 & 0.91 \end{pmatrix}.$$

As with the previous study, we simulate 20150 observations and take the first 150 observations as the training set and the remaining 20000 as the testing set. For the score function, a multivariate Bregmann loss function is used with $\phi(\boldsymbol{x}) = |x_1|^{1.5} + |x_2|^{1.5}$. This form of ϕ generalises the power loss function with $a = 1.5$, see Equation 4.3, and the resulting score function is given by

$$s(\boldsymbol{x}_{t+1}, \boldsymbol{\theta}) = \sum_{i=1}^{2} \left(x_{t+1,i}^{1.5} - \mathbb{E}_{\boldsymbol{\theta}}[X_{t+1,i}]^{1.5} - 1.5\mathbb{E}_{\boldsymbol{\theta}}[X_{t+1,i}]^{0.5}(x_{t+1,i} - \mathbb{E}_{\boldsymbol{\theta}}[X_{t+1,i}]) \right). \tag{8.1}$$

For the penalty function we use the Dahl penalty on the t.p.m; see Equation 5.5. We fit a four-state HMM using Algorithm 1. Ten values of α are considered and the number of starting values is 20. Once again, the cross-validation scheme is Δ-sequential sampling with $\Delta = 3$. As the penalty function is not everywhere differentiable, we use the Nelder-Mead routine to minimise the objective function. The resulting out-of-sample and cross-validation scores are shown in Figure 8.5. As with the previous study, the use of a penalty causes a significant improvement in the forecast accuracy of the HMM. In addition, the cross-validation scheme correctly identifies the value of α minimising the out-of-sample score. For $\alpha > 0.15$, the value of the penalty function can be shown to be approximately zero; that is, the t.p.m. effectively follows the Dahl structure described in Section 5.3.2; we term this a 'structured' HMM. In this case is it clear that the HMM with a structured t.p.m. has a lower out-of-sample score than the unstructured

HMM (see Figure 8.5). However, our proposed approach of basing a penalty upon a particular structure results in a lower out-of-sample score than both the structured and unstructured HMM.

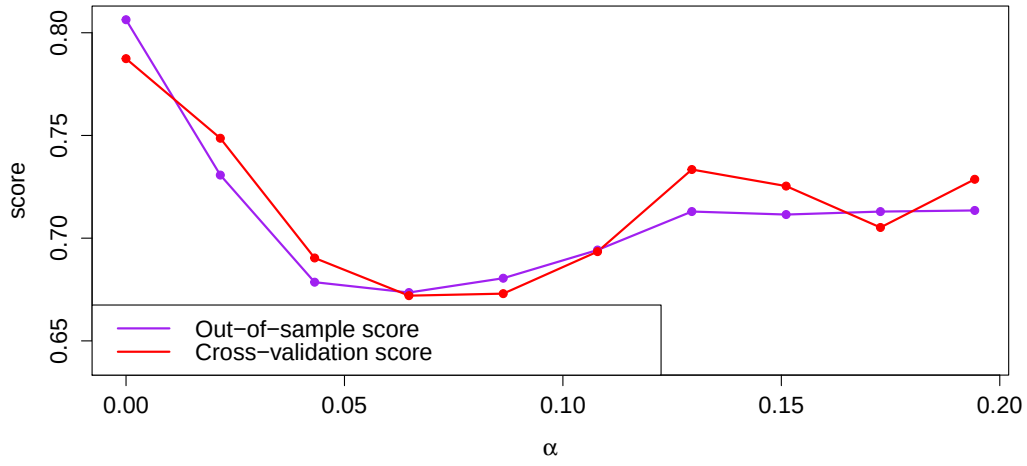


Figure 8.5: Comparison of exponential-HMM's out-of-sample and cross-validation scores for increasing α.

The best HMM identified by the cross-validation scheme is for $\alpha = 0.07$; we term this the 'penalised HMM'. The out-of-sample score of the unpenalised and penalised HMMs is 0.80 and 0.68 respectively. As an aside, we also fit a 4-state HMM using maximum likelihood and then measure the out-of-sample score using (8.1) as 0.92; this demonstrates the benefit of matching the method of parameter estimation with the score function.

As with the previous study, we check both the fit and 95% potential risk for the unpenalised 4-state, penalised and true HMMs; the Q-Q and potential risk plots are shown in Figures 8.6 and 8.7 respectively. In terms of the Q-Q plots, the penalised HMM shows a much better fit to the training data and, critically, the plot for the penalised HMM is similar to the plot for the true HMM. Surprisingly, the unpenalised HMM appears to fit the training data poorly; one would expect the unpenalised model to fit the training

data better than the penalised model. Similarly, the potential risk for the penalised HMM is much lower than that of the unpenalised HMM and, in addition, is very similar to the potential risk for the true HMM.

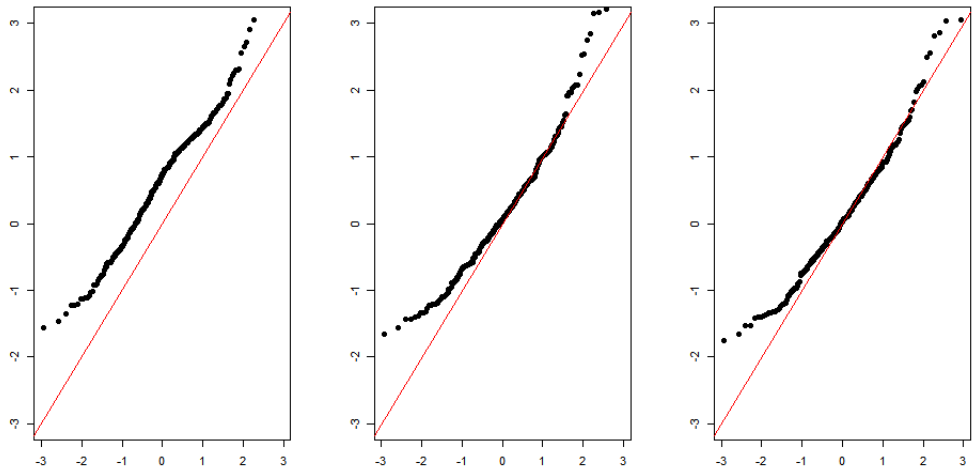


Figure 8.6: Quantile-quantile plots for the unpenalised, penalised and true exponential-HMMs respectively. Theoretical quantiles are shown on the horizontal axis.

This simulation study differed from the previous in terms of the dimension of the observations, the type of HMM fitted, the form of the score function and the choice of penalty. Nonetheless, the conclusion to this simulation study is the same as the previous one; the penalised HMM outperformed the unpenalised HMM in terms of forecast accuracy, model fit and potential risk. Notable in this case was the use of a penalty on the t.p.m. which outperformed both the unpenalised and structured HMM in terms of the out-of-sample score.

8.3 Monthly counts of disability benefit claims

A series of 120 consecutive counts of the number of monthly short-term disability claims made by injured workers to the British Columbia workers'

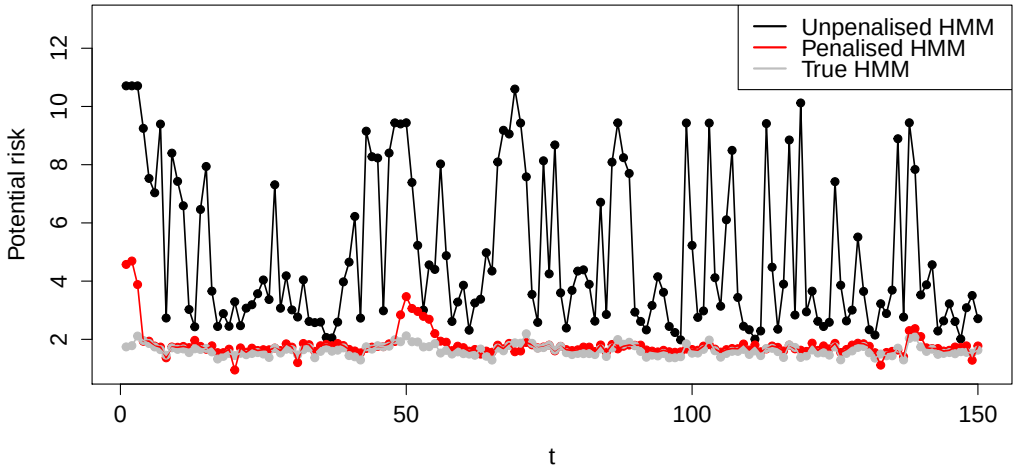


Figure 8.7: Comparison of potential risk for unpenalised, penalised and true categorical-HMM.

compensation board is considered (Freeland, 1998). A Poisson-HMM is a potentially good model for these data as the number of claims per month is effectively unbounded, but the observations are fairly small; the sample mean is 6.13 and the largest observation is 21.

The first 78 observations are taken as the training set and the last 38 observations as the testing set; four observations are removed to reduce dependence between the sets. We follow the approach of Zhu and Joe (2006) by using a squared-error loss function to measure the score of a model fitted on these data. A squared-error loss function is of the Bregmann form and thus

$$s(x_{t+1}, \boldsymbol{\theta}) = (x_{t+1} - \mathbb{E}_{\boldsymbol{\theta}}[X_{t+1}])^2.$$

However, despite using the squared-error loss to measure the error of their proposed model, Zhu and Joe (2006) fit their model using maximum likelihood; we emphasise that Zhu and Joe do not use an HMM, but rather a Markov process based on binomial thinning. Therefore, to begin we demonstrate the advantage of matching the loss function and method of estimation

in terms of the forecast accuracy of the HMM. This is done by fitting the an unpenalised Poisson-HMM twice on the training set; first using a minus log-likelihood score and second using the extremum estimator that follows from the above score function. The optimisation routine used is `nlm`. The forecast accuracy of the models, calculated using the squared-error loss, is then measured using the testing set. We fit the models for two, three and four hidden states. The resulting out-of-sample scores are given in Table 8.1 and show a clear advantage to parameter estimation by minimising the squared-error loss as opposed to the minus log-likelihood. For reference, the out-of-sample error of the 1-state HMM was 9.47. The next step is to attempt to increase

	Score function used to fit model	
Number of states	Minus log-likelihood	Squared-error loss
2	10.52	**8.55**
3	9.59	**7.69**
4	10.11	**8.05**

Table 8.1: Comparison of out-of-sample squared-error loss scores for different estimation techniques and number of hidden states.

the forecast accuracy of the HMM by introducing a penalty function. We propose the use of a KLD penalty on $\boldsymbol{\lambda}$; that is,

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{m} \delta_i \left(\boldsymbol{\delta}'\boldsymbol{\lambda} - \lambda_i + \lambda_i \log \frac{\lambda_i}{\boldsymbol{\delta}'\boldsymbol{\lambda}} \right). \tag{8.2}$$

The above penalty follows from Equation 5.9. Algorithm 1 is then run for 2, 3 and 4 states with 3, 10 and 20 starting values respectively. Ten values of α are considered. The cross-validation scheme used is Δ-sequential sampling with $\Delta = 3$ and `nlm` is used to minimise the objective functions. Surprisingly, the penalty fails to decrease the cross-validation score for the 4-state HMM; thus we consider further the 2 and 3-state HMMs only. In these cases the cross-validation score does decrease; comparative plots of the out-of-sample and cross-validation scores are given in Figures 8.8 and 8.9.

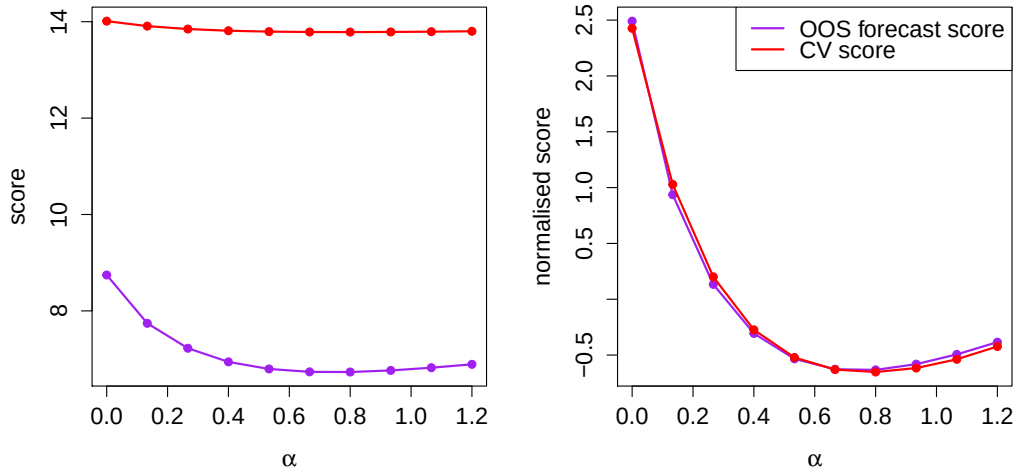


Figure 8.8: Comparison of absolute and normalised out-of-sample and cross-validation scores for 2-state Poisson-HMM.

Observe that the cross-validation scores greatly overestimate the out-of-sample score. A possible reason for this is that the testing set contains far fewer extreme values than the training set; the largest values in the training and testing sets are 21 and 11 respectively. An extreme value is likely to result in a larger score value, especially under squared-error loss.

The large differences between the cross-validation and out-of-sample scores are potentially worrying but, in principle, the key consideration is whether the cross-validation score identifies accurately the value of α minimising the out-of-sample score. Therefore, the normalised out-of-sample and cross-validation scores for the 2 and 3-state HMMs are also given in Figures 8.8 and 8.9 respectively. For the 2-state HMM, the normalised[1] out-of-sample and cross-validation scores are remarkably similar whereas for the 3-state HMM the cross-validation score incorrectly identifies the value of α minimising the out-of-sample score. Nonetheless, based upon the cross-validation scores, we

[1]The normalised scores are calculated by subtracting by the mean and dividing by the standard deviation.

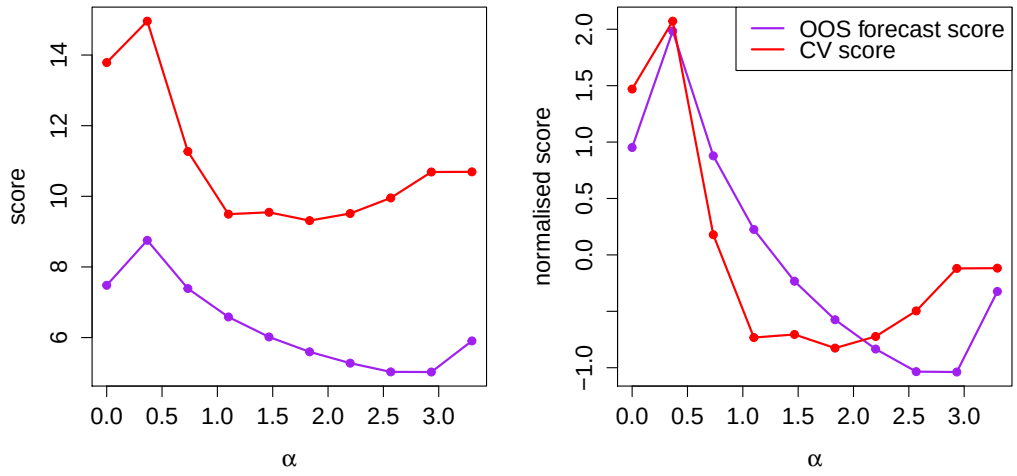


Figure 8.9: Comparison of absolute and normalised out-of-sample and cross-validation scores for 3-state Poisson-HMM.

must take the best possible model as the 3-state HMM with $\alpha = 1.83$; we term this the penalised model. The out-of-sample score of this HMM is 5.60 which is a substantial improvement over any of the unpenalised HMMs. The use of a penalty is further justified by comparing the Q-Q plots and 95% potential risk for the unpenalised and penalised 3-state HMMs; see Figures 8.10 and 8.11. The penalised HMM is a better fit to the training set than the unpenalised HMM. Finally, the penalised HMM has a substantially lower potential risk than the unpenalised HMM.

This study shows the usefulness of both extremum estimators and penalised estimation for HMMs applied to real data. The out-of-sample score for a 3-state unpenalised HMM fitted using a minus log-likelihood score is 9.59 versus 5.60 for a 3-state penalised HMM fitted using the correct score function. Approximately 48% of this decrease can be attributed to matching the method of estimation with the score function of the forecaster, and the remaining 52% to introducing a penalty function.

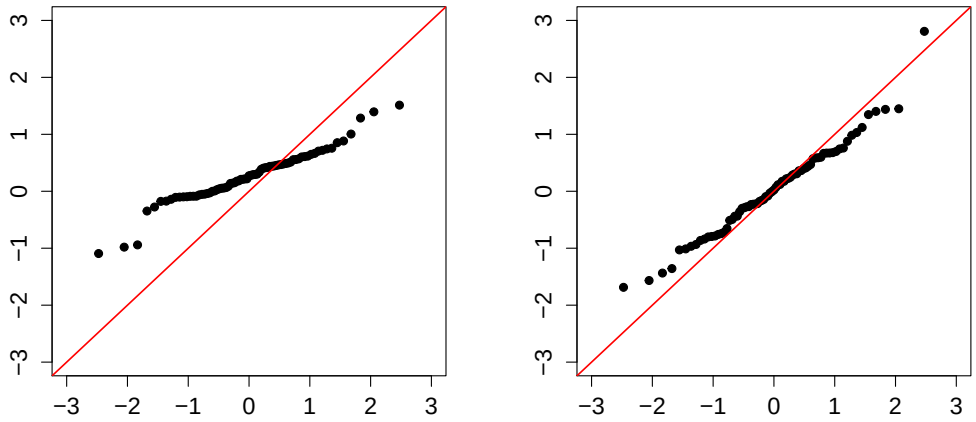


Figure 8.10: Quantile-quantile plots for the unpenalised and penalised 3-state Poisson-HMM respectively. Theoretical quantiles are shown on the horizontal axis.

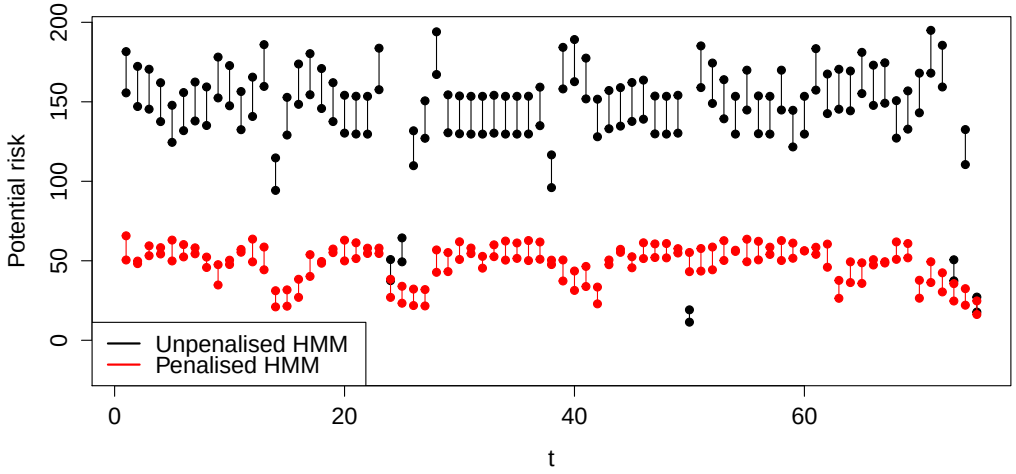


Figure 8.11: Comparison of potential risk intervals for the unpenalised and penalised 3-state Poisson-HMM.

8.4 Daily counts of epileptic seizures

For our final application we consider 204 daily counts of epileptic seizures for one patient (Leroux and Puterman, 1992). We fit a Poisson-HMM, the score function is the squared-error loss and the penalty used is given in Equation 8.2. The first 158 observations are taken as the training set and the last 42 as the testing set; four observations are removed to reduce dependence between the sets.

Algorithm 1 is run for 2 and 3 states. The number of starts and values of α considered is 10. The cross-validation scheme used is last-block validation with the last 30 observations of the training set taken as the validation set. The objective function is minimised using `nlm`. The resulting out-of-sample and cross-validation scores are given in Figure 8.12. For reference, the out-of-sample score of the unpenalised 1-state HMM is 0.80. Once again, the penalty

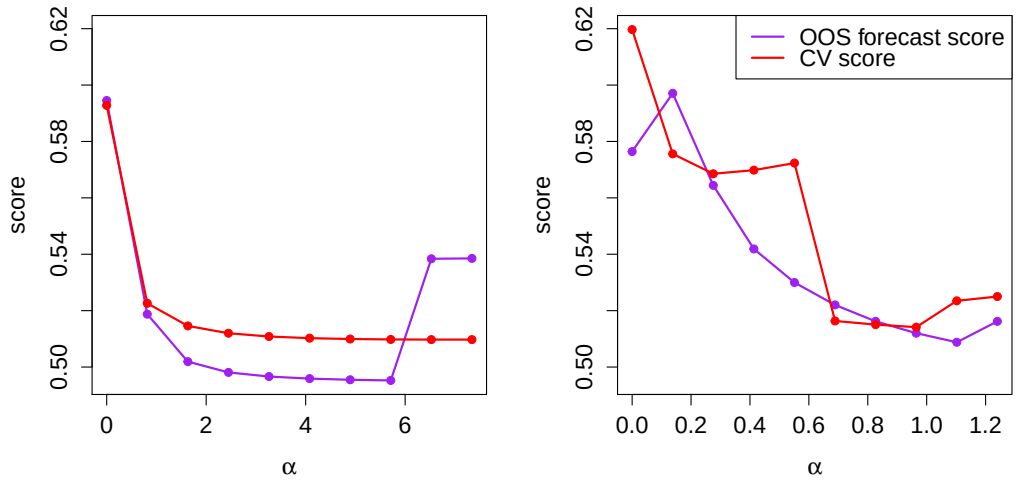


Figure 8.12: Comparison of out-of-sample and cross-validation scores for two and three-state Poisson-HMM respectively.

results in a significant decrease in the out-of-sample score for both the 2 and 3-state HMM and the cross-validation score is a reasonable approximation

to the out-of-sample score. Based upon the cross-validation scores, the best possible HMM is the 2-state HMM with $\alpha = 6.53$; we term this the 'penalised HMM'. Unfortunately, the cross-validation procedure incorrectly identifies the HMM with the lowest possible score which is the 2-state HMM with $\alpha = 5.71$. Nonetheless, the penalised HMM does have a lower out-of-sample score than the unpenalised 2-state HMM.

Next we compare the out-of-sample score for 2-state HMMs fitted using a minus log-likelihood score and a squared-error loss score. Surprisingly, the out-of-sample score for both models is the same and equal to 0.59. However, no improvement to the out-of-sample score could be found by adding a penalty to the HMM fitted using a minus log-likelihood score.

As before, we compare the 2-state unpenalised and penalised HMMs using both Q-Q plots and potential risk; these are given in Figures 8.13 and 8.14 respectively. Both HMMs show a good fit to the training data but the potential risk intervals for the penalised model tend to be lower than those of the unpenalised model. Lower potential risk for the penalised model appears to be a trend throughout the applications; see the comparative potential risks shown in Figures 8.3 and 8.11. An explanation for this is that the penalties on $\boldsymbol{\lambda}$ generally discourage 'extreme' values in the $\boldsymbol{\lambda}$ matrix. Hence, the potential risk for the penalised model is often substantially lower.

As for the previous application, this application demonstrates the usefulness of penalised estimation for HMMs. In this case, the entire improvement in the out-of-sample score is attributable to the introduction of a penalty function. However, the introduction of extremum estimators is still useful as no improvement in the out-of-sample score could be found using penalised likelihood estimation. It is notable in this case that the penalty brings an improvement in the out-of-sample score despite the fairly large training set and $\boldsymbol{\lambda}$ containing only two parameters.

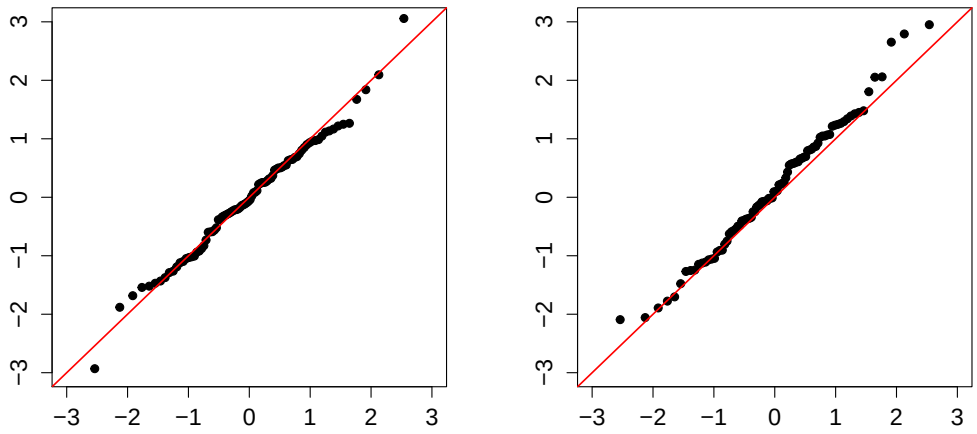


Figure 8.13: Quantile-quantile plots for the unpenalised and penalised 2-state Poisson-HMM respectively. Theoretical quantiles are shown on the horizontal axis.

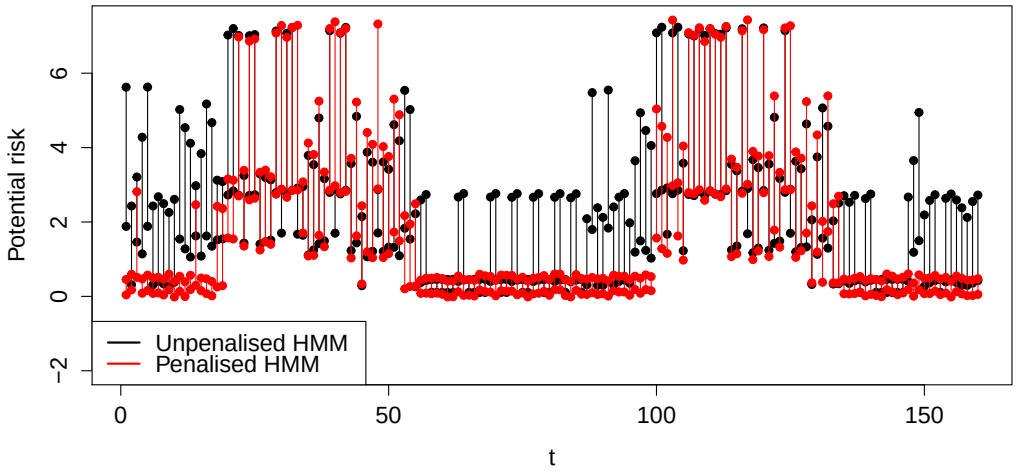


Figure 8.14: Comparison of potential risk intervals for the unpenalised and penalised 2-state Poisson-HMM.

CHAPTER 9

Concluding remarks

The focus of this dissertation has been on improving the forecast accuracy of the HMM. Two major suggestions were made in this regard. First, the application of extremum estimators to HMMs was proposed. This approach allows for consistency between the actual measure of forecast accuracy and the objective function used to fit the model. Second, the addition of a penalty function to the objective function was proposed as a method for increasing the forecast accuracy of an HMM. A number of possible penalty functions were suggested and, in particular, soft structuring was introduced. A cross-validation approach for tuning the penalty function was also described.

A general four-step approach for forecasting with the HMM was introduced. This approach was applied to both simulated and real data using a variety of state-dependent distributions, score functions and penalty functions. These applications demonstrated three key results which support our proposed approach. First, it was shown how matching the measure of forecast accuracy with the method of parameter estimation resulted in improved forecast accuracy. Second, the improvement in forecast accuracy brought by the introduction of a penalty function was demonstrated. Finally, it was shown how cross-validation may be used to tune the penalty function in order to select a tuning parameter value which resulted in improved forecast

accuracy.

Of course, it is acknowledged that the applications presented are limited; the proposed methods may be less successful in a broader range of applications. There also several further areas of research that have become apparent in carrying out this research. We describe three of these areas below.

Penalised estimation was proposed as a method with the single aim of improving the forecast accuracy of the HMM; the precise choice of penalty was not important as long as the accuracy improved. Little consideration was given to how one may determine a good penalty *a priori*, and to examining the properties of the penalties in more detail. Investigating this area may produce useful research.

A basic heuristic optimisation technique was suggested to help overcome the problem of multiple local minima; it seems likely that this technique can be improved. There appears to be at least two possible approaches to this. The first is to consider more sophisticated versions of the general optimisation technique proposed in this dissertation. The second approach is to consider optimisation techniques specific to certain types of HMMs, score functions and penalty functions.

Finally, this dissertation has focused on the basic HMM only; no consideration has been given to the many generalisations of the HMM, for example, hierarchical HMMs (Fine et al., 1998), hidden semi-Markov models (Ferguson, 1980) and HMMs with covariates. These generalisations of the HMM are often less parsimonious than the basic HMM and may therefore be better candidates for penalised estimation. A possible extension would be to investigate the usefulness of penalised estimation for generalisations of the HMM.

APPENDIX A

Consistency of extremum estimators for the HMM

This chapter serves to establish theoretical support for the general class of estimators proposed in Section 3.2. In particular, it is shown that a value of $\boldsymbol{\theta}$ minimising $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$, that is $\widehat{\boldsymbol{\theta}}_T$, will in some sense tend a value of $\boldsymbol{\theta}$ minimising $S_0(\boldsymbol{\theta})$. This property is termed 'consistency', a concept studied extensively in the asymptotic theory of statistics; see, for example, DasGupta (2008).

A.1 Regularity conditions

The analysis presented requires a number of mild regularity conditions to be imposed.

- **Condition 1** (C1). The underlying Markov chain is doubly infinite and hence stationary.
- **Condition 2** (C2). The t.p.m. $\boldsymbol{\Gamma}$ is aperiodic and irreducible.
- **Condition 3** (C3). If Θ is not compact, then for all $i \in M$, $p(\cdot|\lambda_i)$ is locally Lipschitz continuous and $p(x|\cdot)$ is semi-continuous. In addition,

$p(\cdot|\lambda_i) \to 0$ as λ_i tends to the boundary of its parameter space.

A.1.1 Stationarity and ergodicity

C1 is often assumed in the analysis of HMMs, for example, by Leroux and Puterman (1992) and Bickel et al. (1998), in order to establish strict stationarity of the observed process $\{X_t\}_{t\in\mathbb{Z}}$; the definition of which is given below.

Definition 2. *A stochastic process $\{X_t\}_{t\in\mathbb{Z}}$ is **strictly stationary** if the joint distribution of $(X_{t_1}, ..., X_{t_k})$ and $(X_{t_{1+h}}, ..., X_{t_{k+h}})$ is the same for all $t_1, ..., t_k, h \in \mathbb{Z}$ and $k \in \mathbb{N}^+$.*

In the application of HMMs, assuming the hidden process is stationary is also common; see Zucchini and MacDonald (2009) for some examples. Note that this a simplifying assumption in the sense that there are $m - 1$ fewer parameters to be estimated. C1 implies an important result given in Lemma 1.

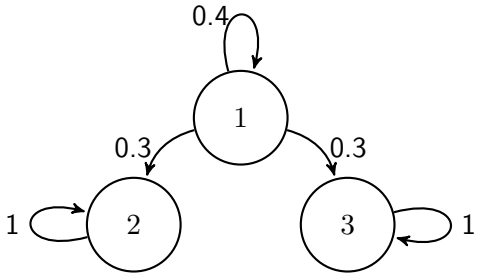
Lemma 1. *If $\{X_t\}_{t\in\mathbb{Z}}$ is an HMM satisfying C1 then it is strictly stationary.*

This property will come in use for studying the asymptotic properties of a general class of estimators in Section A.2. Another important property is ergodicity. The precise definition of ergodicity is beyond the scope of this dissertation; we provide an intuitive explanation below. The key point is that if C2 holds, then the hidden Markov chain $\{C_t\}_{t\in\mathbb{Z}}$ is said to be ergodic, which implies the existence of a unique stationary distribution. The ergodicity of $\{C_t\}_{t\in\mathbb{Z}}$ leads to an important theorem for HMMs.

Lemma 2. *Suppose $\{X_t\}_{t\in\mathbb{Z}}$ is an HMM and C1 and C2 hold. Then $\{X_t\}_{t\in\mathbb{Z}}$ is ergodic.*

Ergodicity and hence strict stationarity help overcome a fundamental problem in basic time-series analysis. That problem being normally only a single sample path of the process over a fixed period is observed but valid

statistical inference requires repeated sampling. If the observed process is assumed strictly stationary, then the time invariance of the joint observation distributions implies multiple observations of such distributions. If in addition the process is assumed ergodic then the process is not too persistent, such that each observation x_t contains some information not available in the other elements. An implication is that a sufficiently long sample of the observed process will be sufficient to obtain valid estimates of the moments of the entire process $\{X_t\}_{t\in\mathbb{Z}}$. To illustrate this point, suppose a non-ergodic three-state HMM has a hidden Markov chain with transition probabilities as described below.



In this case, the observed process $\{X_t\}_{t\in\mathbb{N}}$ will contain draws from at most two of the state-dependent distributions, irrespective of the initial distribution. Thus a valid inference on all the state-dependent distributions cannot be drawn.

A.1.2 Compactness

A technical problem regards the compactness of the parameter space Θ; asymptotic results often require the parameter space to be compact. Recall that a subset of $\mathbb{R}$ is compact if it is closed and bounded. We assume each element of Θ, θ_j, has parameter space $I_j \subseteq \mathbb{R}$ which takes one of the following forms: $(a,b), [a,b), (a,b]$ or $[a,b]$ where $a, b \in \mathbb{R} \cup \{\infty, -\infty\}$. The parameter space Θ, the Cartesian product of each I_j, $\prod_j I_j$, is compact if and only if every I_j is compact. For example, the parameter space for each transition probability $[0,1]$ is compact but the parameter space for the mean of a Poisson

distribution $(0, \infty)$ is not. To resolve the problem of compactness, we adopt the approach of Kiefer and Wolfowitz (1956) and extend each I_j to a compact space I_j^c which denotes the closure of I_j over the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty, -\infty\}$. We also extend each p_i by defining $p_i(x) = 0$ for every 'new' value of x in I_j^c. Then the extended parameter space $\Theta^c = \prod_j I_j^c$ is compact.

To add a concrete example, consider a m-state Poisson HMM where each p_i is parameterised by the mean of p_i, λ_i. Then

$$\Theta = \prod_{j=1}^{m(m-1)} [0,1] \times \prod_{k=1}^{m} (0, \infty)$$

is not compact because each $(0, \infty)$ is both unbounded and open in $\mathbb{R}$. Thus we extend each interval to $[0, \infty]$ which is compact in $\overline{\mathbb{R}}$ and also define $p_i(\infty) = 0$. The new parameter space

$$\Theta^c = \prod_{j=1}^{m(m-1)} [0,1] \times \prod_{k=1}^{m} [0, \infty]$$

is then compact in $\overline{\mathbb{R}}$.

In these cases, C3 ensures that the continuity of each p_i on the extended parameter space. It holds for at least three common state-dependent distributions used in HMMs; the Poisson distribution, Normal distribution with fixed variance and Exponential distribution (Leroux, 1989).

For the sake of simplicity, we shall assume every parameter space Θ has already been made compact, and thus avoid use of the c superscript.

A.2 Establishing consistency of extremum estimators for the HMM

Before moving onto the proof of consistency, a minor technical problem must be overcome. The definition of $S_0(\boldsymbol{\theta})$ in Equation 3.1 is ambiguous in that the expectation depends on a particular value of t; the forecast distribution is a function of the history of the process. This ambiguity emerges from a

similar ambiguity in the definition of forecasting in Section 3.1; the number of previous observations available is not fixed. It will be seen that studying estimation in a rigorous setting requires this value to be specified a priori. For the purposes of this dissertation it is assumed that each forecast is conditioned on the previous T observations, where T is the total of number of observations. The key assumption is that observations that are not available are regarded as missing. It should be emphasised that the introduction of a fixed period aids the mathematics of subsequent results but, as will be seen, has no effect on the value of the estimated parameters.

In demonstrating the consistency of extremum estimators for the HMM, we will require that the score, as a function of the parameter-vector $\boldsymbol{\theta}$, be locally Lipschitz continuous.

Definition 3. *Given two metric spaces (X, d_X) and (Y, d_Y), a mapping $f : X \longrightarrow Y$ is **locally Lipschitz continuous** if for every $x \in X$ there exists a constant $K \in \mathbb{R}_0^+$ and a neighbourhood U of x such that, for all x_1 and x_2 in U,*

$$d_Y(f(x_1), f(x_2)) \leq K d_X(x_1, x_2).$$

Intuitively, a locally Lipschitz continuous function can be thought better behaved than a continuous function in that locally Lipschitz functions are differentiable almost everywhere (Federer, 2014). Thus **Condition 4** (C4) is that $s(\cdot, \boldsymbol{\theta})$ is locally Lipschitz continuous; the metric taken as the L_2 distance.

A.2.1 Preliminary definitions

The following analysis requires first three definitions.

Definition 4. *A sequence of random variables $\{A_t\}_{t\in\mathbb{N}}$ is said to converge **almost surely** (a.s.) to a random variable A if and only if*

$$\Pr\left(\lim_{t\to\infty} A_t = A\right) = 1.$$

An equivalent definition, regarding the random variable A_t as a mapping from a sample space Ω to $\mathbb{R}$ is that $\{A_t\}_{t\in\mathbb{N}}$ converges to a random variable A a.s. if and only if

$$\Pr\left(\omega \in \Omega : \lim_{t\to\infty} A_t(\omega) = A(\omega)\right) = 1.$$

Definition 5. *A sequence of non-negative random variables $\{A_{t,\boldsymbol{\theta}}\}_{t\in\mathbb{N}}$ depending on a parameter $\boldsymbol{\theta} \in \Theta$ is said to converge* ***almost surely uniformly*** *to 0 if and only if*

$$\Pr\left(\lim_{t\to\infty} \sup_{\boldsymbol{\theta}\in\Theta} A_{t,\boldsymbol{\theta}} = 0\right) = 1.$$

As above, an equivalent condition is:

$$\Pr\left(\omega \in \Omega : \lim_{t\to\infty} \sup_{\boldsymbol{\theta}\in\Theta} A_{t,\boldsymbol{\theta}}(\omega) = 0\right) = 1.$$

Definition 6. *A sequence of estimators $\{\widehat{\boldsymbol{\theta}}_T\}_{T\in\mathbb{N}}$ is said to be* ***strongly consistent*** *for a constant parameter $\boldsymbol{\theta}_0$ if and only if $\widehat{\boldsymbol{\theta}}_T$ converges a.s. to $\boldsymbol{\theta}_0$ as $T \to \infty$.*

For the sake of brevity, 'consistency' is henceforth to mean 'strong consistency'. We wish to establish that the estimator $\widehat{\boldsymbol{\theta}}_T$ is consistent where $\boldsymbol{\theta}_0$ is a value minimising the expected score. The literature often refers to $\boldsymbol{\theta}_0$ as the 'true' parameter value in the sense that $\boldsymbol{\theta}_0$ is the assumed parameter value of the actual generating process. Under this interpretation, consistency requires that an estimator, given unlimited data, should reveal the underlying 'truth'. We prefer to avoid this interpretation; for our purposes consistency is purely a property to aid forecasting, an assurance that an estimate tends towards a value providing the best forecasts. The parameter $\boldsymbol{\theta}_0$ should be interpreted precisely and only so far as it is defined mathematically; that is the value which minimises the expected score.

A.2.2 A problem of multiple minima

Unfortunately, Definition 6 is not of direct use in this dissertation; in the context of a non-trivial HMM, both $\arg\min_{\theta\in\Theta} S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ and $\arg\min_{\theta\in\Theta} S_0(\boldsymbol{\theta})$ are sets with cardinality potentially greater than one.

An immediate cause of this problem is label switching, which describes the invariance of the probability of observing a particular series of observations to relabelling of the states in the model. That is, it is possible to reorder the parameter vector $\boldsymbol{\theta}$ without altering the probability measure; $\mathbb{P}_{\boldsymbol{\theta}}$. This problem is well studied in the context of HMMs and can be resolved fairly easy; see, for example, the approach taken by Zucchini and MacDonald (2009).

Unfortunately, even if the label switching problem is resolved, there is no a priori reason to preclude the existence of multiple minima. It is entirely possible that both $S_0(\boldsymbol{\theta})$ and $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ have infinite numbers of minima. For example, compact regions $R \subseteq \Omega$ could arise such that $S_0(\boldsymbol{\theta})$ is minimised for all $\boldsymbol{\theta} \in R$. For many $S_0(\boldsymbol{\theta})$ these 'flat' regions are uncountably infinite and multiple such regions may exist. This problem is perhaps best explained graphically; thus consider a fictitious and very simple univariate $S_0(\theta)$, which is plotted over the region $\Theta = [0, 0.7]$ in Figure A.1. The red regions denote minima, with the red dots showing boundaries of the sets of minima. The single point at 0.1, say $\theta_{0,1}$, is a strict local minimum in the interval $(\theta_{0,1} - \epsilon, \theta_{0,1} + \epsilon)$ for some $\epsilon > 0$. In contrast, no point in the interval $[0.5, 0.6]$ is a strict local minimum. These regions can cause problems when optimising, and are discussed in Section 7. Is is important to accept that, unlike the problem of label switching, both $S_0(\boldsymbol{\theta})$ and $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ have a potentially uncountably infinite number of minima.

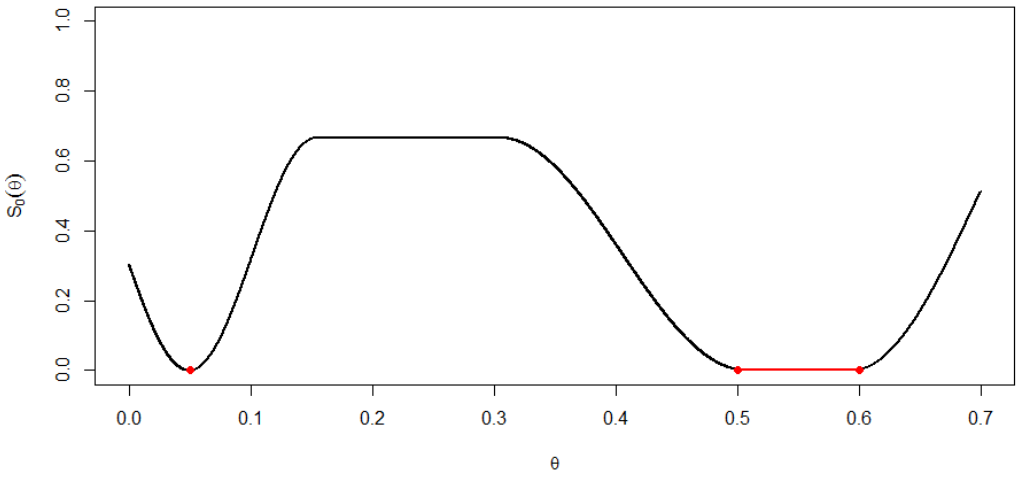


Figure A.1: Minima of $S_0(\theta)$.

It is thus pointless to discuss consistency in terms of a single value of $\widehat{\boldsymbol{\theta}}_T$ and a single value of $\boldsymbol{\theta}_0$. Rather, consistency must be defined in terms of sets of minima. To make explicit the set notation, let

$$\widehat{\Theta}_T = \arg\min_{\boldsymbol{\theta}\in\Theta} S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta}),$$

$$\Theta_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} S_0(\boldsymbol{\theta});$$

those are the sets of parameter vectors minimising the objective function and expected score respectively. A more general definition of consistency in terms of sets is now presented. Required first is a measure of distance between two sets. Thus, for two closed sets A and B in a Euclidean space define the mapping $d_{\mathrm{DH}} : A \times B \longrightarrow \mathbb{R}_0^+$ where

$$d_{\mathrm{DH}}(A, B) = \sup_{a\in A} \inf_{b\in B} \|a - b\|_2,$$

which is termed the directed Hausdorff distance (Deza and Deza, 2009). This measure lends itself to a definition of consistency in terms of sets of estimates.

Definition 7. *A sequence of estimators* $\{\widehat{\boldsymbol{\theta}}_T\}_{T\in\mathbb{N}}$ *with each* $\widehat{\boldsymbol{\theta}}_T \in \widehat{\Theta}_T$ *is said to be* ***strongly consistent*** *for* Θ_0 *if and only if*

$$d_{DH}\left(\widehat{\Theta}_T, \Theta_0\right) = \sup_{\widehat{\boldsymbol{\theta}}_T \in \widehat{\Theta}_T} \inf_{\boldsymbol{\theta}_0 \in \Theta_0} \|\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2$$

converges a.s. to 0 as $T \to \infty$.

This definition has great intuitive appeal: for any $\widehat{\boldsymbol{\theta}}_T \in \widehat{\Theta}_T$ the infimum

$$\inf_{\theta_0 \in \Theta_0} \|\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2$$

identifies the Euclidean distance to the nearest element of Θ_0. The supremum over $\widehat{\Theta}_T$ is thus equal to greatest distance between some element $\widehat{\Theta}_T$ and the nearest element of Θ_0. By requiring this quantity to tend to zero, it is assured with a probability of one that any estimate $\widehat{\boldsymbol{\theta}}_T$ can, for some value of T, be made arbitrarily close to some element of Θ_0. To provided further justification for this choice measure of distance between sets, two other possibilities for the metric used in defining strong consistency are presented. First is the Haussdorf metric:

$$d_{\mathrm{H}}(A, B) = \max\{\sup_{a\in A} \inf_{b\in B} \|a - b\|_2, \sup_{b\in B} \inf_{a\in A} \|a - b\|_2\},$$

see, for example, Menzel (2014), and second the infimum pseudosemimetric

$$d_{\inf}(A, B) = \inf_{a\in A} \inf_{b\in B} \|a - b\|_2,$$

see, for example, Cheng and Liu (2001). Substituting d_{DH} with d_{H} in Definition 7 results in a stronger condition: in addition to every element in $\widehat{\Theta}_T$ approaching Θ_0, every element of Θ_0 must be approached by a sequence in $\widehat{\Theta}_T$. This is a necessary property when one wants to identify the entire set Θ_0 but, when the emphasis is on forecasting, it is sufficient to require the sequence of estimates to approach Θ_0 only; the particular value of $\boldsymbol{\theta}_0$ is of no relevance to the expected score. Substituting d_{DH} with $d_{\inf}$ in Definition 7 results in a weaker condition: it requires the existence of only a single sequence of estimates to approach Θ_0; an insufficient condition for the purposes of this dissertation. Heuristically, it is useful to think of consistency in terms of $\widehat{\Theta}_T$ and Θ_0 when there is an infinite amount of data. Then, for d_{DH} is it required that $\widehat{\Theta}_\infty \subseteq \Theta_0$, for d_{H} that $\widehat{\Theta}_\infty = \Theta_0$ and for $d_{\inf}$ that $\widehat{\Theta}_\infty \cap \Theta_0 \neq \emptyset$.

A.2.3 A proof of consistency

Equipped with Definition 7 it is possible to prove the main theorem of this section. We provide first a required lemma and then state the theorem.

Lemma 3. *If $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \Theta$ and Θ is compact, then $\widehat{\Theta}_T$ and Θ_0 are closed.*

Theorem 3. *Suppose*

1. *Θ is compact,*
2. *$S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and semi-continuous in $\boldsymbol{x}_{1:T}$,*
3. *$|S_T(\boldsymbol{X}_{1:T}, \boldsymbol{\theta}) - S_0(\boldsymbol{\theta})|$ converges to 0 a.s. uniformly in $\boldsymbol{\theta} \in \Theta$ as $T \to \infty$, where $S_0(\boldsymbol{\theta})$ is a non-stochastic function.*

Then $\widehat{\boldsymbol{\theta}}_T$ is strongly consistent in the sense of Definition 7.

Assumptions 1 and 2 are fairly general; the compactness of Θ has already been described and Assumption 2 will hold for most practical applications. The key assumption is the third, which ensures that with probability one that $S_T(\boldsymbol{X}_{1:T}, \boldsymbol{\theta})$ tends to $S_0(\boldsymbol{\theta})$ as T tends to infinity. In fact, the result of Theorem 3 is not surprising given Assumption 3. Unfortunately verifying this assumption can be difficult and it would thus be useful to have a set of assumptions which is sufficient for Assumption 3 of Theorem 3, and generally easier to verify. Such a set does exist, but before stating it the notion of first-moment continuity of a random function is introduced.

Definition 8. *Let*

$$\epsilon_t(\boldsymbol{\theta}, \delta) = \sup\{|s(X_t, \boldsymbol{\theta}) - s(X_t, \boldsymbol{\alpha})| : \boldsymbol{\alpha} \in \Theta \text{ such that } \|\boldsymbol{\alpha} - \boldsymbol{\theta}\|_2 < \delta\}.$$

*Then $s(X_t, \boldsymbol{\theta})$ is **first-moment continuous** at $\boldsymbol{\theta} \in \Theta$ if*

$$\lim_{\delta \to 0} \mathbb{E}[\epsilon_t(\boldsymbol{\theta}, \delta)] = 0,$$

the expectation being taken with respect to X_t.

For the forms imposed here, $S_T(\boldsymbol{x}_{1:T}, \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} s(x_t, \boldsymbol{\theta})$ and $S_0(\boldsymbol{\theta}) = \mathbb{E}[s(X_t, \boldsymbol{\theta})]$, Singleton (2009) gives a set of four assumptions which ensure that Assumption 3 of Theorem 3 holds. We conclude this section by stating this theorem, noting its usefulness in combination with Theorem 3 for demonstrating that a particular form of the score s elicits a consistent estimator.

Theorem 4. *Suppose*

1. Θ *is compact,*
2. $\{X_t\}_{t\in\mathbb{Z}}$ *is ergodic,*
3. $\mathbb{E}[s(X_t, \boldsymbol{\theta})]$ *exists and is finite for all* $\boldsymbol{\theta} \in \Theta$*,*
4. $s(X_t, \boldsymbol{\theta})$ *is first-moment continuous for all* $\boldsymbol{\theta} \in \Theta$*.*

Then

$$\sup_{\boldsymbol{\theta}\in\Theta}\{|S_T(\boldsymbol{X}_{1:T}, \boldsymbol{\theta}) - S_0(\boldsymbol{\theta})|\} \to 0 \text{ a.s. as } T \to \infty.$$

APPENDIX B

Proofs

B.1 General results

Proof of Lemma 1 (See page 89). Provided C1 holds, Zucchini and MacDonald (2009) showed that for any $x_1, ..., x_k \in \Omega$,

$$\Pr(X_{t_1} = x_1, ..., X_{t_k} = x_k) = \boldsymbol{\delta}'\mathbf{P}(x_1)\prod_{j=2}^{k}\boldsymbol{\Gamma}^{t_j - t_{j-1}}\mathbf{P}(x_j)\mathbf{1}$$

for $t_1 < t_2 < ... < t_k$ with $t_1, ..., t_k \in \mathbb{Z}$. Then for any $h \in \mathbb{Z}$,

$$\begin{aligned}
&\Pr(X_{t_1+h} = x_1, ..., X_{t_k+h} = x_k) \\
&= \boldsymbol{\delta}'\mathbf{P}(x_1)\prod_{j=2}^{k}\boldsymbol{\Gamma}^{t_j+h-t_{j-1}-h}\mathbf{P}(x_j)\mathbf{1} \\
&= \Pr(X_{t_1} = x_1, ..., X_{t_k} = x_k),
\end{aligned}$$

and thus $\{X_t\}_{t\in\mathbb{Z}}$ is strictly stationary. □

Proof of Lemma 2 (See page 89). See Leroux and Puterman (1992).

Proof of Theorem 1 (See page 20). See Gneiting (2008).

Proof of Theorem 2 (See page 21). See Gneiting (2008).

B.2 Asymptotic results

Unless stated otherwise, convergence is meant with respect to the L_2 metric.

Proof of Lemma 3 (See page 97). As $S_T(\cdot, \boldsymbol{\theta})$ is a continuous function over a compact set Θ, it follows from the extreme value theorem that $S_T(\cdot, \boldsymbol{\theta})$ achieves at least one minimum on Θ. Thus $\widehat{\Theta}_T \neq \emptyset$ and hence there exists at least one convergent sequence in $\widehat{\Theta}_T$. Denote any such sequence $\{\zeta_n\}$, with limit $c \in \Theta$. We now show that $c \in \widehat{\Theta}_T$.

Let $S_{\min} = \min_{\boldsymbol{\theta} \in \Theta} S_T(\cdot, \boldsymbol{\theta})$, the minimum value of $S_T(\cdot, \boldsymbol{\theta})$ over Θ. Clearly the sequence $\{S_T(\cdot, \zeta_n)\}$ is constant with each element equal to $S_{\min}$ and thus it must be that

$$\lim_{n \to \infty} \{S_T(\cdot, \zeta_n)\} = S_{\min}.$$

But, since $S_T(\cdot, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, it must be that

$$S_T(\cdot, c) = \lim_{n \to \infty} \{S_T(\cdot, \zeta_n)\} = S_{\min},$$

which implies $c \in \widehat{\Theta}_T$.

To see Θ_0 is closed, simply replace $\widehat{\Theta}_T$ with Θ_0 and $S_T(\cdot, \boldsymbol{\theta})$ with $S_0(\boldsymbol{\theta})$ in the above proof. □

The proof below is a generalisation of Theorem 3.2 in Singleton (2009) which allows for multiple minima in the extremum estimator.

Proof of Theorem 3 (See page 97). Before proving the theorem, a slightly more rigorous definition of S_T is required. Thus let $\mathcal{X}$ denote the set of all doubly-infinite realisations of X_t and regard the subscript T as indicating which part of each doubly-infinite realisation should be calculated as being in the score. That is, for $\omega \in \mathcal{X}$,

$$S_T(\omega, \boldsymbol{\theta}) : \mathcal{X} \longrightarrow \mathbb{R}$$

such that

$$S_T(\omega, \boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} s((\omega)_t, \boldsymbol{\theta}).$$

Now, assumption 1 and 2 ensure that Θ_0 is non-empty by the extreme value theorem. Thus we can pick a point $\boldsymbol{\theta}_0$ in Θ_0. A particular sequence of estimators $\{\widehat{\boldsymbol{\theta}}_T\}_{T\in\mathbb{N}}$ is also required. For this purpose first assume the states of the HMM are ordered such that $\|\lambda_1\|_2 < \|\lambda_2\|_2 < ... < \|\lambda_m\|_2$, then let

$$\widehat{\boldsymbol{\theta}}_T = \underset{\boldsymbol{\theta}\in\widehat{\Theta}_T}{\arg\sup} \left\{ \inf_{\boldsymbol{\theta}^*\in\Theta_0} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \right\},$$

which ensures that $\widehat{\boldsymbol{\theta}}_T$ is uniquely defined. Since both $\widehat{\Theta}_T$ and Θ_0 are compact (closed subsets of a compact space, see Lemma 3), it must be that $\widehat{\boldsymbol{\theta}}_T \in \widehat{\Theta}_T$.

Now define the function

$$\rho(\epsilon) = \inf\{S_0(\boldsymbol{\theta}) - S_0(\boldsymbol{\theta}_0) : \inf_{\boldsymbol{\theta}^*\in\Theta_0} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \geq \epsilon\}.$$

If $\epsilon > 0$, then $\boldsymbol{\theta}$ is well separated from Θ_0 and following from Conditions 1 and 2 of the theorem and the definition of Θ_0, it must be that $\rho(\epsilon) > 0$. Condition 3 of theorem implies the existence, with probability one, of a function $T(\omega, \epsilon)$ such that

$$\rho_T(\omega) = \sup_{\boldsymbol{\theta}\in\Theta} |S_T(\omega, \boldsymbol{\theta}) - S_0(\boldsymbol{\theta})| < \rho(\epsilon)/2,$$

for all $\omega \in \mathcal{X}$, $\epsilon > 0$ and $T > T(\omega, \epsilon)$. Thus for all $\omega \in \mathcal{X}$, $\epsilon > 0$ and $T > T(\omega, \epsilon)$:

$$\begin{aligned}
S_0(\widehat{\boldsymbol{\theta}}_T) - S_0(\boldsymbol{\theta}_0) =& S_0(\widehat{\boldsymbol{\theta}}_T) - S_T(\omega, \widehat{\boldsymbol{\theta}}_T) + S_T(\omega, \widehat{\boldsymbol{\theta}}_T) \\
&- S_T(\omega, \boldsymbol{\theta}_0) + S_T(\omega, \boldsymbol{\theta}_0) - S_0(\boldsymbol{\theta}_0) \\
\leq& S_0(\widehat{\boldsymbol{\theta}}_T) - S_T(\omega, \widehat{\boldsymbol{\theta}}_T) + S_T(\omega, \boldsymbol{\theta}_0) - S_0(\boldsymbol{\theta}_0) \\
\leq& |S_0(\widehat{\boldsymbol{\theta}}_T) - S_T(\omega, \widehat{\boldsymbol{\theta}}_T)| + |S_T(\omega, \boldsymbol{\theta}_0) - S_0(\boldsymbol{\theta}_0)| \\
\leq& 2\rho_T(\omega) < \rho(\epsilon).
\end{aligned}$$

The first inequality follows from the definition of S_T; $S_T(\omega, \widehat{\boldsymbol{\theta}}_T) \leq S_T(\omega, \boldsymbol{\theta}_0)$. The second inequality follows from the triangle inequality and the remainder of the proof from the definition of $\rho_T(\omega)$.

Thus it has been show that, with probability one, for all $\epsilon > 0$ and $T > T(\omega, \epsilon)$,

$$S_0(\widehat{\boldsymbol{\theta}}_T) - S_0(\boldsymbol{\theta}_0) < \inf\{S_0(\boldsymbol{\theta}) - S_0(\boldsymbol{\theta}_0) : \inf_{\boldsymbol{\theta}^*\in\Theta_0} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \geq \epsilon\},$$

implying $\inf_{\boldsymbol{\theta}^* \in \Theta_0} \|\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|_2 < \epsilon$ under the same conditions. But this is equivalent to

$$\inf_{\boldsymbol{\theta}^* \in \Theta_0} \|\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|_2 \to 0 \text{ a.s. as } T \to \infty,$$

and since by definition

$$\inf_{\boldsymbol{\theta}^* \in \Theta_0} \|\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|_2 = \sup_{\boldsymbol{\theta} \in \widehat{\Theta}_T} \inf_{\boldsymbol{\theta}^* \in \Theta_0} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2,$$

we conclude that $d_{\mathrm{DH}}(\widehat{\Theta}_T, \Theta_0) \to 0$ a.s. as $T \to \infty$. □

B.3 Penalised estimators

This section proves the forms of the penalty function for the t.p.m. structures described in Section 5.3.2. In call cases the premetric d is taken to be that specified in Section 5.3.2, and $\boldsymbol{\Gamma}^{(s)}$ denotes the t.p.m. associated with the parameter vector $\boldsymbol{\theta}_s$; and similarly for $\boldsymbol{\theta}$ and $\boldsymbol{\Gamma}$.

Lemma 4. *Let Θ_{tri} be the parameter space implied by the restrictions of the $\boldsymbol{\Gamma}_{tri}$ structure. Then for any $\boldsymbol{\theta}$,*

$$\min_{\boldsymbol{\theta}_s \in \Theta_{tri}} d(\boldsymbol{\theta}, \boldsymbol{\theta}_s) = \sum_{i,j \in M : |i-j|>1} \gamma_{ij} = J(\boldsymbol{\theta}).$$

It follows that J is a penalty function.

Proof. Clearly $J(\boldsymbol{\theta}_s) = 0$ if $\boldsymbol{\theta}_s \in \Theta_{\mathrm{tri}}$ and $J(\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta} \in \Theta$. Now suppose $\boldsymbol{\theta} \in \Theta$, then

$$\begin{aligned}
d(\boldsymbol{\theta}, \boldsymbol{\theta}_s) &= \sum_{i,j \in M : |i-j|>1} |\gamma_{ij} - \gamma_{ij}^{(s)}| + \sum_{i,j \in M : |i-j| \leq 1} |\gamma_{ij} - \gamma_{ij}^{(s)}| \\
&= \frac{1}{2} J(\boldsymbol{\theta}) + \sum_{i,j \in M : |i-j| \leq 1} |\gamma_{ij} - \gamma_{ij}^{(s)}| \\
&\geq J(\boldsymbol{\theta}) + \left| \sum_{i,j \in M : |i-j| \leq 1} \gamma_{ij} - \sum_{i,j \in M : |i-j| \leq 1} \gamma_{ij}^{(s)} \right| \\
&= J(\boldsymbol{\theta}) + |m - \frac{1}{2} J(\boldsymbol{\theta}) - m| = J(\boldsymbol{\theta}).
\end{aligned}$$

To see this lower bound is attained, take $\gamma_{ij}^{(s)} \geq \gamma_{ij}$ for $i, j \in M : |i-j| \leq 1$. Then $\mathbf{\Gamma}^{(s)}$ is certainly a t.p.m. and

$$\begin{aligned} d(\boldsymbol{\theta}, \boldsymbol{\theta}_s) &= \frac{1}{2} J(\boldsymbol{\theta}) + \sum_{i,j \in M : |i-j| \leq 1} \left(\gamma_{ij}^{(s)} - \gamma_{ij} \right) \\ &= \frac{1}{2} J(\boldsymbol{\theta}) + \sum_{i,j \in M : |i-j| \leq 1} \gamma_{ij}^{(s)} - \left(m - \frac{1}{2} J(\boldsymbol{\theta}) \right) = J(\boldsymbol{\theta}). \end{aligned}$$

Thus we conclude that

$$J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}_s \in \Theta_{\text{tri}}} d(\boldsymbol{\theta}, \boldsymbol{\theta}_s)$$

is a penalty function. □

Lemma 5. *Let $\Theta_{d.s.}$ be the parameter space implied by the restrictions of the $\mathbf{\Gamma}_{d.s.}$ structure. Then for any $\boldsymbol{\theta}$,*

$$\min_{\boldsymbol{\theta}_s \in \Theta_{d.s.}} d(\boldsymbol{\theta}, \boldsymbol{\theta}_s) = \sum_{j=1}^{m} \left| \sum_{i=1}^{m} \gamma_{ij} - 1 \right| = J(\boldsymbol{\theta}).$$

It follows that J is a penalty function.

Proof. Clearly $J(\boldsymbol{\theta}_s) = 0$ if $\boldsymbol{\theta}_s \in \Theta_{\text{d.s.}}$ and $J(\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta} \in \Theta$. Now suppose $\boldsymbol{\theta} \in \Theta$, then

$$\begin{aligned} d(\boldsymbol{\theta}, \boldsymbol{\theta}_s) &= \sum_{j=1}^{m} \sum_{i=1}^{m} |\gamma_{ij} - \gamma_{ij}^{(s)}| \\ &\geq \sum_{j=1}^{m} \left| \sum_{i=1}^{m} \gamma_{ij} - \sum_{i=1}^{m} \gamma_{ij}^{(s)} \right| \\ &= \sum_{j=1}^{m} \left| \sum_{i=1}^{m} \gamma_{ij} - 1 \right| = J(\boldsymbol{\theta}). \end{aligned}$$

To see this lower bound is attained, construct $\mathbf{\Gamma}^{(s)}$ such that, for all i,

$$\begin{aligned} \gamma_{ij}^{(s)} &\geq \gamma_{ij} \quad \text{if } \textstyle\sum_{i=1}^{m} \gamma_{ij} \leq 1 \\ \gamma_{ij}^{(s)} &\leq \gamma_{ij} \quad \text{if } \textstyle\sum_{i=1}^{m} \gamma_{ij} \geq 1. \end{aligned}$$

This can be achieved my increasing or decreasing the transition probabilities in each row until a doubly-stochastic t.p.m. is attained. It follows then that

$$\sum_{j=1}^{m}\sum_{i=1}^{m}|\gamma_{ij}-\gamma_{ij}^{(s)}|=\sum_{j=1}^{m}\left|\sum_{i=1}^{m}\gamma_{ij}-\sum_{i=1}^{m}\gamma_{ij}^{(s)}\right|,$$

and thus the lower bound is attained. We conclude that

$$J(\boldsymbol{\theta})=\min_{\boldsymbol{\theta}_s\in\Theta_{\text{d.s.}}}d(\boldsymbol{\theta},\boldsymbol{\theta}_s)$$

is a penalty function. □

Lemma 6. *Let Θ_{Dahl} be the parameter space implied by the restrictions of the $\boldsymbol{\Gamma}_{Dahl}$ structure. Define $\widehat{\boldsymbol{\gamma}}\in[0,0.5]^{3m-2}$ such that*

$$(\widehat{\boldsymbol{\gamma}})_i=\begin{cases}\max\{1-\gamma_{ii},0.5\} & \text{for } i=1,m,\\ \max\{(1-\gamma_{ii})/2,0.5\} & \text{for } i=2,...,m-1,\\ \max\{\gamma_{i-m,i+1-m},0.5\} & \text{for } i=m+1,...,2m-1,\\ \max\{\gamma_{i-2m+1,i-2m},0.5\} & \text{for } i=2m,...,3m-2,\end{cases}\tag{B.1}$$

and let $\widehat{\gamma}=\min\{\widehat{\boldsymbol{\gamma}}_{1/2},0.5\}$. Then for any $\boldsymbol{\theta}$,

$$\begin{aligned}\min_{\boldsymbol{\theta}_s\in\Theta_{Dahl}}d(\boldsymbol{\theta},\boldsymbol{\theta}_s)&=|\gamma_{11}+\widehat{\gamma}-1|+|\gamma_{mm}+\widehat{\gamma}-1|+\sum_{i=2}^{m-1}|\gamma_{ii}+2\widehat{\gamma}-1|\\&\quad+\sum_{i,j\in M:|i-j|=1}|\gamma_{ij}-\widehat{\gamma}|\\&=J(\boldsymbol{\theta}).\end{aligned}$$

It follows that J is a penalty function.

Proof. Suppose first that $\boldsymbol{\theta}\in\Theta_{\text{Dahl}}$. Then clearly $\widehat{\gamma}=\gamma$ for every i, and thus

$J(\boldsymbol{\theta}) = 0$. In addition, $J(\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta} \in \Theta$. Now suppose $\boldsymbol{\theta} \in \Theta$, then

$$\begin{aligned}
d(\boldsymbol{\theta}, \boldsymbol{\theta}_s) &= \sum_{j=1}^{m} \sum_{i=1}^{m} |\gamma_{ij} - \gamma_{ij}^{(s)}| \\
&= |\gamma_{11} - (1-\gamma)| + |\gamma_{mm} - (1-\gamma)| + \sum_{i=2}^{m-1} |\gamma_{ii} - (1-2\gamma)| \\
&+ \sum_{i=1}^{m-1} (|\gamma_{i,i+1} - \gamma| + |\gamma_{i+1,i} - \gamma|) + \sum_{|i-j|>1} \gamma_{ij} \\
&= |(\widehat{\boldsymbol{\gamma}})_i - \gamma| + \sum_{i=2}^{m-1} |(\widehat{\boldsymbol{\gamma}})_i - \gamma| + |(\widehat{\boldsymbol{\gamma}})_m - \gamma| + \sum_{i=m+1}^{3m-2} |(\widehat{\boldsymbol{\gamma}})_i - \gamma| + \sum_{|i-j|>1} \gamma_{ij} \\
&= \sum_{i=1}^{3m-2} |(\widehat{\boldsymbol{\gamma}})_i - \gamma| + \sum_{|i-j|>1} \gamma_{ij}.
\end{aligned}$$

The value of γ minimising the last term is the median of $\widehat{\boldsymbol{\gamma}}$ or, if this median is greater than 0.5, the term is minimised by $\gamma = 0.5$. This is precisely the definition of $\widehat{\gamma}$ and it follows that J is a penalty function. □

References

B. Abdous and R. Theodorescu. Note on the spatial quantile of a random vector. *Statistics & Probability Letters*, 13(4):333–336, 1992.

T. Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *Information Theory, IEEE Transactions on*, 51(7):2664–2669, 2005.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300, 1995.

C. Bergmeir and J. M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

P. J. Bickel, Y. Ritov, and T. Rydén. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4):1614–1635, 1998.

P. J. Bickel, B. Li, A. B. Tsybakov, S. A. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, and A. van der Vaart. Regularization in statistics. *Test*, 15(2):271–344, 2006.

C. G. Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

J. Bulla and I. Bulla. Structured hidden Markov models. Computing in Economics and Finance 2006 437, Society for Computational Economics, 2006. URL `http://EconPapers.repec.org/RePEc:sce:scecfa:437`.

P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3): 503–514, 1989.

P. Burman, E. Chow, and D. Nolan. A cross-validatory method for dependent data. *Biometrika*, 81(2):351–358, 1994.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2002.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

O. E. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics, 2005.

G. Celeux and J.-B. Durand. Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics*, 23(4):541–564, 2008.

R. Cheng and W. Liu. The consistency of estimators in finite mixture models. *Scandinavian Journal of Statistics*, 28(4):603–616, 2001.

F. H. Clarke. A new approach to Lagrange multipliers. *Mathematics of Operations Research*, 1(2):165–174, 1976.

B. Cooper and M. Lipsitch. The analysis of hospital infection data using hidden Markov models. *Biostatistics*, 5(2):223–237, 2004.

G. Dahl. Tridiagonal doubly stochastic matrices. *Linear Algebra and its Applications*, 390:197–208, 2004.

A. DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Science & Business Media, 2008.

A. P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147:278–292, 1984.

J. De Leeuw and W. J. Heiser. Convergence of correction matrix algorithms for multidimensional scaling. *Geometric Representations of Relational Data*, pages 735–752, 1977.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 39:1–38, 1977.

M. M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, 2009.

F. X. Diebold, T. A. Gunther, and A. S. Tay. Evaluating density forecasts: With applications to financial risk management. *International Economic Review*, 39:863–883, 1998.

M. N. Do. Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models. *Signal Processing Letters, IEEE*, 10(4): 115–118, 2003.

D. Duffie and J. Pan. An overview of value at risk. *The Journal of Derivatives*, 4(3):7–49, 1997.

P. K. Dunn and G. K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

J. E. Ebel, D. W. Chambers, A. L. Kafka, and J. A. Baglivo. Non-poissonian earthquake clustering and the hidden markov model as bases for earthquake forecasting in california. *Seismological Research Letters*, 78(1):57–65, 2007.

B. Efron and R. Tibshirani. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92 (438):548–560, 1997.

H. Federer. *Geometric Measure Theory*. Springer, 2014.

J. Ferguson. Variable duration models for speech. In *Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech*, pages 143–179. Princeton, New Jersey, 1980.

S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden Markov model: analysis and applications. *Machine learning*, 32(1):41–62, 1998.

R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, 1970.

R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149–154, 1964.

R. K. Freeland. *Statistical analysis of discrete time series with application to the analysis of workers' compensation claims data*. PhD thesis, University of British Columbia, 1998.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

T. Gneiting. Quantiles as optimal point predictors. *University of Washington, Technical Report*, (538), 2008.

T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

A. M. González, A. Roque, and J. García-González. Modeling and forecasting electricity prices with input/output hidden Markov models. *Power Systems, IEEE Transactions on*, 20(1):13–24, 2005.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

M. R. Hassan and B. Nath. Stock market forecasting using hidden Markov model: a new approach. In *Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on*, pages 192–196. IEEE, 2005.

T. Hastie, J. Friedman, and R. Tibshirani. *The Elements of Statistical Learning*. Springer, New York, 2009.

F. Hayashi. Econometrics. 2000. *Princeton University Press. Section*, 1: 60–69, 2000.

J. U. Hjorth. *Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap*. CRC Press, 1993.

A. E. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58(3):54–59, 1962.

P. J. Huber. *Robust Statistics*. Springer, Berlin Heidelberg, 2011.

R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

W. Karush. Minima of functions of several variables with inequalities as side constraints. Master's thesis, University of Chicago, 1939.

B. Keller and R. Lutz. Improved learning for hidden Markov models using penalized training. In *Artificial Intelligence and Cognitive Science*, pages 53–60. Springer, 2002.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Annals of Mathematical Statistics*, 27(4):887–906, 1956.

H. Kuhn and A. Tucker. Nonlinear programming. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 481–492, 1951.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

R. Langrock, I. L. MacDonald, and W. Zucchini. Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models. *Journal of Empirical Finance*, 19(1):147–161, 2012.

H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on machine learning*, pages 473–480. ACM, 2007.

B. G. Leroux. Maximum-likelihood estimation for mixture distributions and hidden Markov models. *cIRcle: UBC's Digital Repository: Electronic Theses and Dissertations (ETDs)*, 1989. URL `http://hdl.handle.net/2429/29176`.

B. G. Leroux and M. L. Puterman. Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, 48: 545–558, 1992.

Z.-H. Ling and L.-R. Dai. Minimum Kullback–Leibler divergence parameter generation for HMM-based speech synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(5):1492–1502, 2012.

I. L. MacDonald. Numerical maximisation of likelihood: A neglected alternative to EM? *International Statistical Review*, 82(2):296–308, 2014.

R. T. McGibbon, B. Ramsundar, M. M. Sultan, G. Kiss, and V. S. Pande. Understanding protein dynamics with L_1-regularized reversible hidden Markov models. *arXiv preprint arXiv:1405.1444*, 2014.

K. Menzel. Consistent estimation with many moment inequalities. *Journal of Econometrics*, 182(2):329–350, 2014.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

K. Osband and S. Reichelstein. Information-eliciting compensation schemes. *Journal of Public Economics*, 27(1):107–115, 1985.

N. D. Pham. *Improved Nelder Meads simplex method and applications*. PhD thesis, Auburn University, 2012.

B. T. Polyak. *Introduction to Optimization*. Optimization Software New York, 1987.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

J. Racine. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1):39–61, 2000.

A. W. Robertson, S. Kirshner, and P. Smyth. Downscaling of daily rainfall occurrence over Northeast Brazil using a hidden Markov model. *Journal of Climate*, 17(22):4407–4424, 2004.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

A. P. Ruszczyński. *Nonlinear Optimization*, volume 13. Princeton University Press, 2006.

R. Schlaifer and H. Raiffa. *Applied Statistical Decision Theory.* Harvard University Press, 1961.

R. Serfling. Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, 56(2):214–232, 2002.

D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

K. J. Singleton. *Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment.* Princeton University Press, 2009.

C. G. Small. A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, 58:263–277, 1990.

N. Städler and S. Mukherjee. Penalized estimation in high-dimensional hidden Markov models with state-specific graphical models. *The Annals of Applied Statistics*, 7(4):2157–2179, 2013.

L. J. Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838, 1992.

V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Science & Business Media, 2000.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1:80–83, 1945.

M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, 1968.

R. Zhu and H. Joe. Modelling count data time series with Markov processes based on binomial thinning. *Journal of Time Series Analysis*, 27(5):725–738, 2006.

W. Zucchini and P. Guttorp. A hidden Markov model for space-time precipitation. *Water Resources Research*, 27(8):1917–1923, 1991.

W. Zucchini and I. L. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R*, volume 110. Chapman and Hall/CRC, 2009.